



JPW

ASA-1162

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of

A. SATOYAMA et al

Serial No. 10/766,823

Filed: January 30, 2004

For: STORAGE SYSTEM AND REPLICATION CREATION METHOD THEREOF

TRANSMITTAL OF CERTIFIED PRIORITY DOCUMENT

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Submitted herewith is a certified priority document
(JP 2003-403868, filed December 3, 2003) of a corresponding
Japanese patent application for the purpose of claiming
foreign priority under 35 U.S.C. §119. An indication that
this document has been safely received would be appreciated.

Respectfully submitted,

Daniel J. Stanger
Registration No. 32,846
Attorney for Applicants

MATTINGLY, STANGER & MALUR
1800 Diagonal Rd., Suite 370
Alexandria, Virginia 22314
(703) 684-1120
Date: June 22, 2004

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 2003年12月 3日
Date of Application:

出願番号 特願2003-403868
Application Number:
[ST. 10/C]: [JP 2003-403868]

出願人 株式会社日立製作所
Applicant(s):

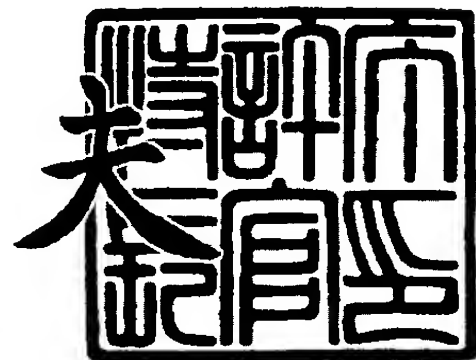
A. Satoyama et al
10/766,823
filed 1-30-04
703-684-1120
ASA-1162



2004年 1月27日

特許庁長官
Commissioner,
Japan Patent Office

今井 康夫



出証番号 出証特2004-3003015

【書類名】 特許願
【整理番号】 K03010501
【あて先】 特許庁長官殿
【国際特許分類】 G06F 12/00
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 里山 愛
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 山本 康友
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 森下 昇
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 江口 賢哲
【特許出願人】
 【識別番号】 000005108
 【氏名又は名称】 株式会社日立製作所
【代理人】
 【識別番号】 100083552
 【弁理士】
 【氏名又は名称】 秋田 収喜
 【電話番号】 03-3893-6221
【手数料の表示】
 【予納台帳番号】 014579
 【納付金額】 21,000円
【提出物件の目録】
 【物件名】 特許請求の範囲 1
 【物件名】 明細書 1
 【物件名】 図面 1
 【物件名】 要約書 1

【書類名】 特許請求の範囲**【請求項 1】**

複数のディスク装置が接続された制御部を複数個有する記憶装置システムであって、
前記制御部に接続されたディスク装置内のボリュームのデータのレプリケーションを作成するレプリケーション作成部と、

レプリケーション元のボリュームとレプリケーション先のボリュームに関する情報であるペア情報と、

を各制御部毎に備え、

前記複数の制御部のうちの一制御部のレプリケーション作成部は、

同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は、前記ペア情報に、レプリケーション元のボリューム情報と、レプリケーション先のボリューム情報と、を登録し、前記ペア情報に基づいて前記同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成し、

他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は、前記ペア情報に、レプリケーション元のボリューム情報と、前記一制御部におけるレプリケーション先のボリューム情報と、前記他の制御部に関する情報と、を登録し、前記ペア情報に基づいて、レプリケーションを作成するための要求を前記他の制御部へ送信することによりレプリケーションを作成することを特徴とする記憶装置システム。

【請求項 2】

請求項 1 に記載の記憶装置システムであって、

前記各制御部はデータを一時的に記憶するキャッシュを備え、

前記一制御部のレプリケーション作成部は、

他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合に、レプリケーション元のボリュームのデータを、前記一制御部における前記キャッシュ内にレプリケーション先のボリュームのデータ用にコピーし、前記キャッシュ内にコピーされた前記データを、前記他の制御部に送信することを特徴とする記憶装置システム。

【請求項 3】

請求項 2 に記載の記憶装置システムであって、

前記他の制御部は、前記一制御部から受信した前記データを、該制御部のキャッシュに格納した後に、該制御部に接続されたディスク装置内のボリュームに格納することを特徴とする記憶装置システム。

【請求項 4】

請求項 1 に記載の記憶装置システムであって、

前記一制御部におけるレプリケーション先のボリューム情報は、前記一制御部に接続されたディスク装置にボリュームが作成されることがない仮想的なボリューム情報であり、

前記他の制御部に関する情報は、前記他の制御部を識別するための情報と前記他の制御部に接続されたディスク装置内のレプリケーション先のボリュームのボリューム情報である

ことを特徴とする記憶装置システム。

【請求項 5】

請求項 1 に記載の記憶装置システムであって、

前記ペア情報は、レプリケーション元のボリュームとレプリケーション先のボリュームのペアに対して割り当てられた識別子を含み、

一個のレプリケーション元のボリュームに対して一個または複数の識別子を組むことを特徴とする記憶装置システム。

【請求項 6】

請求項 1 に記載の記憶システムであって、

前記レプリケーション作成部は、

管理端末から入力された情報若しくはホストからのホストコマンドに基づいて、または、自動的に、
前記ペア情報への情報の登録を行うことを特徴とする記憶装置システム。

【請求項 7】

請求項 1 記載の記憶装置システムであって、
通常のリード／ライト要求をスケジュールするための通常リード／ライト処理キューと、
通常のリード／ライト要求よりも処理の優先度が低い要求をスケジュールするための処理優先度低キューと、
を各制御部毎に備え、
前記レプリケーションを作成するための要求を通常リード／ライト要求と同等に処理する場合は、該要求を前記通常リード／ライト処理キューに入れ、
前記レプリケーションを作成するための要求を通常リード／ライト要求より低い優先度で処理する場合は、該要求を前記処理優先度低キューに入れる
ことを特徴とする記憶装置システム。

【請求項 8】

請求項 1 に記載の記憶装置システムであって、
前記レプリケーションを作成するための要求を通常のリード／ライト要求と同等に処理すべきかどうかの処理優先度を設定し、他の制御部へ処理優先度情報を通知する処理優先度設定部を各制御部毎に備え、
前記通知を受けた他の制御部が、前記処理優先度情報により、要求を処理する順番をスケジュールすることを特徴とする記憶装置システム。

【請求項 9】

請求項 1 に記載の記憶装置システムであって、
一制御部が、他の制御部へレプリケーションを作成するための要求を送信する場合に、送信する制御命令にレプリケーション作成処理であることを示す情報を付加して送信し、
前記他の制御部が、前記情報に基づいて前記要求を優先処理するかどうかを決定し、前記要求を処理する順番をスケジュールすることを特徴とする記憶装置システム。

【請求項 1 0】

請求項 1 に記載の記憶装置システムであって、
一制御部が、他の制御部へレプリケーションを作成するための要求を送信する場合に、送信する要求命令に優先度を示す情報を付加して送信し、
前記他の制御部が、前記情報に基づいて前記要求を優先処理するかどうかを決定し、前記要求を処理する順番をスケジュールすることを特徴とする記憶装置システム。

【請求項 1 1】

複数のディスク装置が接続された制御部を複数個有する記憶装置システムにおけるレプリケーション作成方法であって、
前記記憶装置システムは、
前記制御部に接続されたディスク装置内のボリュームのデータのレプリケーションを作成するレプリケーション作成部と、
レプリケーション元のボリュームとレプリケーション先のボリュームに関する情報であるペア情報と、
を各制御部毎に備え、
同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は、前記複数の制御部のうちの一制御部のレプリケーション作成部が、前記ペア情報に、レプリケーション元のボリューム情報と、レプリケーション先のボリューム情報と、を登録するステップと、
他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は、前記一制御部のレプリケーション作成部が、前記ペア情報に、レプリケーション元のボリューム情報と、前記一制御部におけるレプリケーション先のボリューム情報と、前

記他の制御部に関する情報と、を登録するステップと、

同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は、前記一制御部のレプリケーション作成部が、前記ペア情報に基づいて前記同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成するステップと、

他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は、前記一制御部のレプリケーション作成部が、前記ペア情報に基づいて、レプリケーションを作成するための要求を前記他の制御部へ送信することによりレプリケーションを作成するステップと

を有することを特徴とするレプリケーション作成方法。

【請求項 1 2】

請求項 1 1 に記載のレプリケーション作成方法であって、

前記記憶装置システムの前記各制御部はデータを一時的に記憶するキャッシュを備え、

他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合に、前記一制御部のレプリケーション作成部が、

レプリケーション元のボリュームのデータを、前記一制御部における前記キャッシュ内にレプリケーション先のボリュームのデータ用にコピーし、前記キャッシュ内にコピーされた前記データを、前記他の制御部に送信する

ことを特徴とするレプリケーション作成方法。

【請求項 1 3】

請求項 1 2 に記載のレプリケーション作成方法であって、

前記他の制御部は、前記一制御部から受信した前記データを、該制御部のキャッシュに格納した後に、該制御部に接続されたディスク装置内のボリュームに格納することを特徴とするレプリケーション作成方法。

【請求項 1 4】

請求項 1 1 に記載のレプリケーション作成方法であって、

前記一制御部におけるレプリケーション先のボリューム情報は、前記一制御部に接続されたディスク装置にボリュームが作成されることがない仮想的なボリューム情報であり、

前記他の制御部に関する情報は、前記他の制御部を識別するための情報と前記他の制御部に接続されたディスク装置内のレプリケーション先のボリュームのボリューム情報である

ことを特徴とするレプリケーション作成方法。

【請求項 1 5】

請求項 1 1 に記載のレプリケーション作成方法であって、

前記ペア情報は、レプリケーション元のボリュームとレプリケーション先のボリュームのペアに対して割り当てられた識別子を含み、

一個のレプリケーション元のボリュームに対して一個または複数の識別子を組むことを特徴とするレプリケーション作成方法。

【請求項 1 6】

請求項 1 1 に記載のレプリケーション作成方法であって、

前記レプリケーション作成部は、

管理端末から入力された情報若しくはホストからのホストコマンドに基づいて、または、自動的に、

前記ペア情報への情報の登録を行うことを特徴とするレプリケーション作成方法。

【請求項 1 7】

請求項 1 1 に記載のレプリケーション作成方法であって、

前記記憶装置システムは、

通常のリード／ライト要求をスケジュールするための通常リード／ライト処理キューと

、通常のリード／ライト要求よりも処理の優先度が低い要求をスケジュールするための処理優先度低キューと、

を各制御部毎に備え、

前記レプリケーションを作成するための要求を通常リード／ライト要求と同等に処理する場合は、該要求を前記通常リード／ライト処理キューに入れ、

前記レプリケーションを作成するための要求を通常リード／ライト要求より低い優先度で処理する場合は、該要求を前記処理優先度低キューに入れることを特徴とするレプリケーション方法。

【請求項 1 8】

請求項 1 1 に記載のレプリケーション作成方法であって、

前記記憶装置システムは処理優先度設定部を各制御部毎に備え、

一制御部の処理優先度設定部が、前記レプリケーションを作成するための要求を通常のリード／ライト要求と同等に処理すべきかどうかの処理優先度を設定し、他の制御部へ処理優先度情報を通知し、

前記通知を受けた他の制御部が、前記処理優先度情報により、要求を処理する順番をスケジュールすることを特徴とするレプリケーション作成方法。

【請求項 1 9】

請求項 1 1 に記載のレプリケーション作成方法であって、

一制御部が、他の制御部へレプリケーションを作成するための要求を送信する場合に、送信する制御命令にレプリケーション作成処理であることを示す情報を付加して送信し、

前記他の制御部が、前記情報に基づいて前記要求を優先処理するかどうかを決定し、前記要求を処理する順番をスケジュールすることを特徴とするレプリケーション作成方法。

【請求項 2 0】

請求項 1 1 に記載のレプリケーション作成方法であって、

一制御部が、他の制御部へレプリケーションを作成するための要求を送信する場合に、送信する要求命令に優先度を示す情報を付加して送信し、

前記他の制御部が、前記情報に基づいて前記要求を優先処理するかどうかを決定し、前記要求を処理する順番をスケジュールすることを特徴とするレプリケーション作成方法。

【書類名】 明細書

【発明の名称】 記憶装置システムおよびそのレプリケーション作成方法

【技術分野】

【 0 0 0 1 】

本発明は、複数のディスク装置が接続された制御部を複数個有する記憶装置システムおよびそのレプリケーション作成方法に関する。

【背景技術】

【 0 0 0 2 】

近年、企業が保有する記憶装置に格納されたデータの複製を、別の記憶装置に作成する処理（以下、「バックアップ」と称する）に係る時間を短縮したいという要求が高まっている。これは、企業が保有する情報量の増加に伴ってバックアップに係る時間が益々増加する一方で、企業の業務時間の延長により、バックアップの処理に割り当てられる時間が短くなっていることが背景にある。

【 0 0 0 3 】

特許文献 1、特許文献 2 に開示されたように、企業における日常の業務を停止せずに、記憶装置に格納されたデータをバックアップする技術として、スナップショットが提案されている。スナップショットとは、記憶装置と接続される計算機を介さずに、記憶装置が有する記憶領域の特定の時点におけるコピーを記憶装置に作成する機能である。この機能を利用して、ユーザは、元の記憶領域を業務で使用し、コピーされた記憶領域に格納されたデータをバックアップに使用する。

【 0 0 0 4 】

【特許文献 1】 特開平 7 - 2 1 0 4 3 9 号公報

【特許文献 2】 特開 2 0 0 1 - 3 1 8 8 3 3 号公報

【発明の開示】

【発明が解決しようとする課題】

【 0 0 0 5 】

ネットワークに接続される記憶装置のスケラビリティを高めるための技術として、クラスタ構成記憶装置システムが考えられる。クラスタ構成記憶装置システムとは、ディスクアレイ装置のような従来の記憶装置システムを 1 クラスタとし、一つの記憶装置システムを複数のクラスタから構成した記憶装置システムである。

【 0 0 0 6 】

従来、クラスタ構成記憶装置システムにおいてスナップショットを行うことを示唆した文献はない。また、クラスタ構成記憶装置システムと従来のスナップショット技術を単純に組み合わせた場合、1 クラスタ内に限って、記憶領域のコピーを実行する技術となる。

【 0 0 0 7 】

しかし、異なるクラスタ間で記憶領域のコピーを作成することが出来ないと、1 つのクラスタ構成記憶装置システム内で、記憶領域のコピー先として使用できる記憶領域と使用できない記憶領域ができてしまい、クラスタ構成記憶装置システムの本来の目的であるスケラビリティが損なわれる。

【 0 0 0 8 】

また、クラスタ構成記憶装置システムにおいて、クラスタ間にまたがった論理ボリューム（以下、「ボリューム」という。）のコピーを作成する場合、つまり、コピー元ボリュームとコピー先ボリュームが異なるクラスタにある場合、コピー元のボリュームがあるクラスタ（以下正クラスタと呼ぶ）は、コピー先のボリュームがあるクラスタ（以下副クラスタと呼ぶ）内の共有メモリを参照できないため、副クラスタの負荷状態について認識できない。このため、コピー元と同一クラスタ内でコピー先のボリュームを選択するという、ユーザの制約が生じるため、従来と装置構成が異なることによるユーザの使い勝手が変更になる。

【 0 0 0 9 】

本発明の目的は、複数のディスク装置が接続された制御部を複数個有する記憶装置シス

テムにおいて、同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合だけでなく、異なった制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合も、異なった制御部であることを意識することなく記憶領域のコピーを作成することができる記憶装置システムを提供できるようにすることである。

【課題を解決するための手段】

【0 0 1 0】

前記目的を達成するために、本発明の記憶装置システムは、複数のディスク装置が接続された制御部を複数個有する記憶装置システムであって、前記制御部に接続されたディスク装置内のボリュームのデータのレプリケーションを作成するレプリケーション作成部と、レプリケーション元のボリュームとレプリケーション先のボリュームに関する情報であるペア情報と、を各制御部毎に備える。そして、前記複数の制御部のうちの一制御部のレプリケーション作成部は、同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は、前記ペア情報に、レプリケーション元のボリューム情報と、レプリケーション先のボリューム情報と、を登録し、前記ペア情報に基づいて前記同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成し、他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は、前記ペア情報に、レプリケーション元のボリューム情報と、前記一制御部におけるレプリケーション先のボリューム情報と、前記他の制御部に関する情報と、を登録し、前記ペア情報に基づいて、レプリケーションを作成するための要求を前記他の制御部へ送信することによりレプリケーションを作成する。

【発明の効果】

【0 0 1 1】

本発明により、複数のディスク装置が接続された制御部を複数個有する記憶装置システムにおいて、同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合だけでなく、異なった制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合も、異なった制御部であることを意識することなく記憶領域のコピーを作成することができる。

【発明を実施するための最良の形態】

【0 0 1 2】

図 1 に本発明の一実施例のクラスタ構成記憶装置システムを含む計算機システムの一例を示す。

第一の記憶装置システム 7 0 A は、ネットワーク 6 0 を介してホスト 1 0, 1 1 と接続されている。管理端末 8 0 は、管理用ネットワーク 6 2 を介して、第二の記憶装置システム 7 0 B は、ネットワーク 6 1 を介して、第一の記憶装置システム 7 0 A に接続される。ここでは、第二の記憶装置システム 7 0 B が接続されている構成を図示するが、7 0 B が接続されていない構成でもよい。

【0 0 1 3】

記憶装置システム 7 0 A は、プロトコル変換アダプタ 4 0 を有する。プロトコル変換アダプタ 4 0 は、記憶制御装置 2 0 と独立したチャンネル接続部分であり、L A N (Local Area Network)、公衆回線、専用回線、E S C O N (Enterprise Systems Connection) に準拠したプロトコルなどを扱う。複数のプロトコル変換アダプタ 4 0 と複数の記憶制御装置 2 0 はネットワーク 6 3 を介して接続されている。

【0 0 1 4】

プロトコル変換アダプタ 4 0 は、ホスト 1 0, 1 1 から入出力要求を受け取ると、コマンドを解析してプロトコル変換し、前記コマンドが要求するデータが格納されている L U (Logical Unit) がどの記憶制御装置 2 0 の配下に管理されているか、記憶装置システム 7 0 B に管理されているかを判断し、判断した箇所に当該要求を送信する。前記どの装置で管理されているかは、ネットワーク 6 3 で接続されたプロセッサ 5 0 内のメモリ上に格納された構成情報テーブル 5 1 0 を参照することで判断する。

【0 0 1 5】

管理端末 8 0 は、ネットワーク 6 2 を介して、第一の記憶装置システム 7 0 A を認識する。また、管理端末 8 0 を、第一の記憶装置システム 7 0 A と専用線で直接接続する構成もある。

【0 0 1 6】

記憶制御装置 2 0 は、CPU 2 1、メモリ 2 2、ホスト 1 0、1 1 等からの入出力データを一時的に格納するキャッシュ 2 3、ネットワーク 6 3 と接続する部分であるハブ 2 4、記憶装置群 3 1 との間のデータの送受信を制御する記憶装置 I/F 2 5 を有し、これらは、内部バスで相互に接続される。

【0 0 1 7】

図 2 に、メモリ 2 2 の一実施例の構成図を示す。メモリ 2 2 には、図 2 に示すように、CPU 2 1 で実行される種々のプログラムが格納される。具体的には、記憶装置システム 7 0 の動作を制御するための RAID 制御プログラム 2 0 0 及び記憶装置システム 7 0 の構成を管理するための管理エージェント 2 1 0 である。又、メモリ 2 2 には、各種管理情報も格納される。具体的には、データのコピー元（レプリケーション元）とコピー先（レプリケーション先）に関する情報を記録するペア情報テーブル 2 2 0、ボリューム情報テーブル 2 3 0、処理優先度情報ビットマップ 2 5 0、差分ビットマップ 2 4 0、処理キュー 2 6 0、記憶装置システム 7 0 B が記憶装置システム 7 0 A に対し自身の LU を記憶装置システム 7 0 A の LU として提供するための構成情報テーブル（図示せず）を有する。前記 2 2 0 ~ 2 5 0 及び構成情報テーブルはプロセッサ 5 0 においてもよい。

【0 0 1 8】

RAID (Redundant Array of Inexpensive Disks) 制御プログラム 2 0 0 は、記憶装置群 3 1 にコマンドを発行する部分（不図示）を有する。RAID 制御プログラム 2 0 0 内には、サブプログラムとして、記憶装置システム 7 0 内でデータのレプリケーション（コピー）を作成するためのレプリケーション作成プログラム 2 0 1 と処理優先度設定プログラム 2 0 2 がある。RAID 制御プログラム 2 0 0、レプリケーション作成プログラム 2 0 1、処理優先度設定プログラム 2 0 2 は CPU 2 1（図 1 参照）で実行され、それにより、記憶制御装置 2 0 の RAID 制御部、レプリケーション作成部、処理優先度設定部が構成される。また、RAID 制御プログラム 2 0 0、レプリケーション作成プログラム 2 0 1、処理優先度設定プログラム 2 0 2 の一部または全部をハードウェアで構成してもよい。尚、データのレプリケーションを実施する際、同期（データのコピーの完了を待って、上位装置に完了を報告する）、非同期（データのコピーの完了を待たずに、上位装置に完了を報告する）といったバリエーションがあるが、本実施形態では、特に区別しない。

管理エージェント 2 1 0 は、管理端末 8 0 からの入力を受けて、情報を設定したり、情報を管理端末 8 0 へ出力するためのプログラムである。

【0 0 1 9】

図 3 に、管理端末 8 0 の一実施例の構成図を示す。管理端末 8 0 は、CPU 8 1、主記憶 8 2、入力部（キーボード装置等）8 3、出力部（ディスプレイ装置等）8 4、ネットワーク 6 2 と接続するための管理 I/F 8 5 及び、記憶部 8 6 を有し、これらは、内部バスで相互に接続される。記憶部 8 6 には、CPU 8 1 で実行されるプログラムである処理優先度設定プログラム 2 0 2 が格納される。

ホスト 1 0 は、例えば、パソコンやワークステーション、汎用コンピュータなどであり、外部に接続するための FC インタフェースである HBA (Host Bus Adaptor)（不図示）を備える。HBA にも WWN (World Wide Name) が付与されている。

【0 0 2 0】

図 4 は、ペア情報テーブル 2 2 0 の一例を示す図である。本テーブルは、記憶装置システム 7 0 内の、コピーされたデータを保持するボリュームの対（以下「ペア」）を管理するための情報であり、ペア番号 2 2 1、正ボリューム情報 2 2 2、副ボリューム情報 2 2 3 ~ 2 2 5、ペア状態 2 2 6 がある。特に、記憶制御装置 2 0 A のペア及び記憶制御装置

2 0 A にコピー元があるペアの情報は、記憶制御装置 2 0 A 内のメモリ 2 2 におく。

ペア番号 2 2 1 は、ペアに対して任意に割り当てられた識別子を示す。

正ボリューム情報 2 2 2 は、識別子が与えられたペアのうち、レプリケーション元（コピー元）のボリュームである正ボリュームに割り振られたボリューム番号を示す。

【0 0 2 1】

副ボリューム情報 2 2 3 ～ 2 2 5 は、識別子が与えられたペアのうち、レプリケーション先（コピー先）のボリュームである副ボリュームに関する情報である。同一記憶制御装置内のペアの場合は、副ボリューム番号を 2 2 3 に登録する。異記憶制御装置間のペアの場合は、同一記憶制御装置内に仮想化したボリューム番号を 2 2 3 に登録し、実際にデータの格納される副ボリュームの記憶制御装置に関する情報、例えば記憶制御装置番号を 2 2 4 に、ボリューム番号を 2 2 5 に登録する。同一記憶制御装置内に仮想化したボリューム番号は、その同一記憶制御装置におけるボリューム番号であるが、その同一記憶制御装置に接続されたディスク装置にボリュームが作成されることがない仮想的なボリューム番号である。

【0 0 2 2】

ペア状態 2 2 6 は、ペアの現在の状態を示す。例えば、ペアの各々のボリュームに格納されているデータの同期が取られ、各々に格納されたデータの内容が一致している状態（以下「P a i r 状態」と称する）又は、ペア間でデータの同期が取られていない状態（以下「S p l i t 状態」と称する）などがある。

【0 0 2 3】

記憶装置システム 7 0 A は、例えば、P a i r 状態にあったペアを、任意の時間において S p l i t 状態に変更する。このとき、副ボリュームには、任意の時間におけるペアの正ボリュームが有するデータが保存される（このような処理を「スナップショットを取る」という）。その後、ホスト 1 0 が副ボリュームからデータを読み出して他の記憶装置（例えばテープ装置等）に書き込むことで、スナップショットを取った時間におけるペアに格納されたデータのバックアップをとることができる。尚、スナップショットを取った後の副ボリュームそのものをデータのバックアップとして保存しておいてもよい。

【0 0 2 4】

図 1 0 は、ボリューム情報テーブル 2 3 0 の一例を示す図である。本テーブルは、記憶制御装置 2 0 A 配下のボリュームを管理する情報を登録し、記憶制御装置 2 0 A 内のメモリに格納する。

ボリューム番号 2 3 1 は、ボリュームに対して割り当てられた識別子である。ボリューム番号 0 は、ボリューム番号 0 に対してペアを 3 個組んだ例である。一般には、1 個のボリュームに対して 1 個または複数個のペアを組むことができる。

正／副 2 3 2 は、ボリュームがペアの正の役割をしているか、副の役割をしているかを示す。

【0 0 2 5】

相手ボリューム情報の 2 3 3 ～ 2 3 5 は、ペアを組んでいる場合の相手のボリューム情報である。同一記憶制御装置内のペアの場合は、副ボリューム番号を 2 3 3 に登録する。異記憶制御装置間のペアの場合は、同一記憶制御装置内に仮想化したボリューム番号を 2 3 3 に登録し、実際にデータの格納される副ボリュームの記憶制御装置番号を 2 3 4 に、ボリューム番号を 2 3 5 に登録する。ボリューム使用中 2 3 6 は、そのボリュームが使用中か空いているかを示す情報である。

【0 0 2 6】

図 1 0 は記憶制御装置 0 番のボリューム情報であるとする。ボリューム 0 番は 3 つのペアを作成している。第 1 のペアは、記憶制御装置 0 番のボリューム 1 0 2 4 番を仮想化副ボリュームとし、実際のデータの格納されている副ボリュームは記憶制御装置 1 番のボリューム 2 0 番であることを示す。第 2 のペアは同一記憶制御装置内のボリューム 1 5 8 番であることを示す。ボリューム 1 番は、ペアの副ボリュームとして使用されており、正ボリュームは記憶制御装置 3 番のボリューム 3 7 8 3 番であることを示す。

【0 0 2 7】

図 1 1 は、差分ビットマップ 2 4 0 の一例を示す図である。1 つのペアに対して同じサイズのビットマップを 2 つ用意する。差分ビットマップの '0' はコピーが終了した箇所、'1' はコピーが終了していない箇所を表している。1 ビットに対して予め決めたデータサイズのデータを対応させている。例えば、1 ビットに対し 6 4 K B のデータを対応させた場合、6 4 K B 中の 1 B でも更新があった場合は、ビットを '1' にして、内容がコピー先にも反映されるようにする。

【0 0 2 8】

図 1 4 は、処理優先度情報ビットマップ 2 5 0 の一例を示す図である。本ビットマップは、ビットマップが格納されている記憶制御装置配下のボリュームで、他の記憶制御装置配下のボリュームとレプリケーションのペアを組んでいるものに対する情報を示す。ビットが '1' の場合、通常のリード／ライト処理より優先度を下げて処理したいことを示す。ビットが '0' の場合は、通常のリード／ライト処理と同等にスケジュールして処理したいことを示す。図 1 4 のビットマップはボリューム番号順に対応しているとする、先頭のビットはボリューム番号 0 の情報となる。このビットマップ例のうち、ビットが '1' であるのは、4 番目、5 番目、8 番目であることから、ボリューム番号 3、4、7 の処理は通常リード／ライト処理の要求頻度が低い時に処理して、通常リード／ライト処理の性能に影響を与えないように処理することを示す。本ビットマップ例では '0' と '1' で示したが、優先度を数段階にわけてスケジュールさせたい場合、1 ボリュームに対するビット数を増やしても良い。

【0 0 2 9】

このシステム構成において、記憶制御装置 2 0 A にあるボリュームのコピーボリュームを記憶制御装置 2 0 B に作る場合の方式について説明する。

コピーボリュームの作成処理は、レプリケーション作成プログラム 2 0 1 が実施する。レプリケーション作成プログラム 2 0 1 は、レプリケーションペアの正副ボリュームが同じ記憶制御装置内か、異なる記憶制御装置間かをチェックし、異記憶制御装置間の場合は、本発明の処理を行う。同一記憶制御装置内のペアの場合は、従来の処理を行う。

【0 0 3 0】

図 5 に、図 1 のクラスタ構成記憶装置システムでレプリケーション作成する際のレプリケーション作成処理の概要のフローチャートを示す。まず、空きボリュームの中からコピー先となる副ボリュームを選択し、正副ボリュームをペアとして、ペア情報テーブル 2 2 0 に登録する（ステップ 5 0 1 0）。ペアが同一記憶制御装置 2 0 内にあるか否かを判定し（ステップ 5 0 2 0）、同一記憶制御装置 2 0 内にある場合は、同一記憶制御装置内のレプリケーション作成処理を行い（ステップ 5 0 3 0）、処理を終了する。もし、異なる記憶制御装置 2 0 間である場合は、異記憶制御装置間レプリケーション作成処理を行い（ステップ 5 0 4 0）、処理を終了する。

【0 0 3 1】

次に、前記ステップ 5 0 3 0 の同一記憶制御装置内レプリケーション作成処理方式を図 6 のフローチャートに従って説明する。

正ボリュームの内容を副ボリュームに全コピーする初期コピー処理を開始する。初期コピーは、図 2、図 1 1 に示す差分ビットマップ 2 4 0 P 1 の全ビットを '1' にする（ステップ 6 0 1 0）。処理が順次差分ビットマップ上の '1' を検出したら（ステップ 6 0 2 0）、そのビットに対応する箇所のデータがキャッシュにあるか否かの判定を行う（ステップ 6 0 3 0）。ステップ 6 0 2 0 で検出しなければ、ステップ 6 0 7 0 へ進む。ステップ 6 0 3 0 でキャッシュになれば正ボリュームからキャッシュにリードする（ステップ 6 0 4 0）。キャッシュ内で副ボリュームのデータ用にコピーする（ステップ 6 0 5 0）。コピーに伴い、データが正しいか否かを判定するための冗長情報も副ボリューム用に新規に作成しデータへ付随させる。キャッシュに格納したら、差分ビットを '0' にする（ステップ 6 0 6 0）。次のビットがあれば、ステップ 6 0 2 0 から 6 0 6 0 を繰り返す（ステップ 6 0 7 0）。次のビットが無ければ、本処理は終了する。また、前記の正ボリ

ュームからデータをリードする際に、副ボリューム用の冗長情報を作成し直接副ボリューム用のデータとしてキャッシュに格納する方法もある。一方、非同期でキャッシュ上の副ボリューム用データを副ボリュームに格納する（ステップ 6 0 8 0）。

【 0 0 3 2 】

次に、前記ステップ 5 0 4 0 の異記憶制御装置間レプリケーション作成処理方法を図 7、図 8 を用いて説明する。

図 7 に、異記憶制御装置間レプリケーションを作成する一実施例を示す。本例では、ボリューム 3 1 1 をコピー元としボリューム 3 1 2 にコピーしたい。その際、正側ボリューム 3 1 1 のある記憶制御装置 2 0 A にて、記憶制御装置 2 0 B にある副側ボリューム 3 1 2 を記憶制御装置 2 0 A 内のボリューム 3 1 3 に仮想化する。これにより、ホストは同一記憶制御装置 2 0 A 内のペア（ボリューム 3 1 1 と 3 1 3 がペアでペア番号 1）としてレプリケーション作成処理を実行する。一方、記憶制御装置 2 0 A は、実際に副ボリュームは記憶制御装置 2 0 B 内にあることを認識し、コピーデータを記憶制御装置 2 0 B へ送信するためのライト要求を発行する。ペア番号 1（ペア # 1）のペア情報は、正ボリュームのある記憶制御装置 2 0 A 内のペア情報テーブル 2 2 0 に登録し、ペア番号 2（ペア # 2）のペア情報は、記憶制御装置 2 0 B 内のペアであるため、記憶制御装置 2 0 B 内のペア情報テーブル 2 2 0 に登録する。

【 0 0 3 3 】

レプリケーションを作成する場合、同一記憶制御装置内で作成することが望ましく、なるべく同一記憶制御装置内で作成するようにするが、レプリケーション元のボリュームがある記憶制御装置に集中したりした場合、同一記憶制御装置内に空き領域がなくなり、複数ある記憶制御装置間でレプリケーションを作成する場合が発生する。このような場合に異記憶制御装置間レプリケーションを作成する一実施例が図 7 に示すものであり、ここでは、正ボリューム 3 1 1 のレプリケーションを実副ボリューム 3 1 2 にとっている。この場合でも、正ボリューム 3 1 1 のレプリケーションを仮想副ボリューム 3 1 3 にとるようにレプリケーション先を選択することにより、同一記憶制御装置内で作成しているように見せかける。仮想副ボリューム 3 1 3 は、ディスク装置内のデータ領域は使わず、ボリューム番号だけ使っている。これが仮想的に割り付ける処理である。ホスト計算機は、正ボリューム 3 1 1 のレプリケーションを仮想副ボリューム 3 1 3 にとるように認識しているため、レプリケーション先がどの記憶制御装置かを認識せずに従来どおりに処理できる。ホスト計算機は、正ボリューム 3 1 1 と仮想副ボリューム 3 1 3 のある記憶制御装置に要求を出せばよい。一方、記憶装置システム 7 0 側では仮想副ボリューム 3 1 3 は仮想的に割り付けられたボリュームであることと、実際のボリュームは実副ボリューム 3 1 2 であることを認識している。このため、実副ボリューム 3 1 2 への要求がきたら、仮想副ボリューム 3 1 3 の記憶制御装置で受け、実副ボリューム 3 1 2 のある記憶制御装置へ要求内容を送信することになる。

【 0 0 3 4 】

図 8 は、異記憶制御装置間レプリケーション作成処理方式を示すフローチャート例である。前記ステップ 5 0 4 0 の異記憶制御装置間レプリケーション作成処理方式を図 8 のフローチャートに従って説明する。

【 0 0 3 5 】

ステップ 8 0 1 0 ～ 8 0 7 0 は、ステップ 6 0 1 0 ～ 6 0 7 0 と同じである。図 8 には詳細なステップは図示していないが、次のように処理が行われる。ステップ 8 0 1 0（ステップ 6 0 1 0 に対応）で図 2、図 1 1 に示す差分ビットマップ 2 4 0 P 1 の全ビットを‘1’とし、ステップ 8 0 2 0（ステップ 6 0 2 0 に対応）で差分ビットマップ上の‘1’を検出したら、ステップ 8 0 3 0（ステップ 6 0 3 0 に対応）でそのビットに対応する箇所のデータがキャッシュにあるか否かの判定を行い、ステップ 8 0 2 0 で‘1’を検出しなければ、ステップ 8 0 7 0（ステップ 6 0 7 0 に対応）へ進む、ステップ 8 0 3 0 でキャッシュになればステップ 8 0 4 0（ステップ 6 0 4 0 に対応）で正ボリュームからキャッシュにリードし、ステップ 8 0 5 0（ステップ 6 0 5 0 に対応）で仮想化したキャ

ッシュ内で副ボリュームのデータ用にコピーし、キャッシュに格納したら、ステップ 8 0 6 0 (ステップ 6 0 6 0 に対応) で差分ビットを '0' にし、ステップ 8 0 7 0 (ステップ 6 0 7 0 に対応) で、次のビットがあれば、ステップ 8 0 2 0 から 8 0 6 0 を繰り返し、次のビットが無ければ、終了する。このようにして、キャッシュ上に仮想化した副ボリューム用データが格納される。ここまでの処理は、正副ボリュームが同じ記憶制御装置内にある場合と同様である。

【0 0 3 6】

一方、非同期でキャッシュ上の仮想化した副ボリューム用データを実際の副ボリュームに格納する必要がある。そのために、正側記憶制御装置 2 0 A からレプリケーションを作成するための要求を副側記憶制御装置 2 0 B へ送信し、副側記憶制御装置 2 0 B 内にレプリケーションを作成する。具体的には、次のようにして、副側記憶制御装置 2 0 B に接続されたディスク装置にレプリケーションを作成する。正ボリュームのある記憶制御装置 2 0 A は、副ボリュームを作成する記憶制御装置 2 0 B の記憶装置に反映されていないキャッシュ上のデータ (ダーティデータ) をライトする要求を記憶制御装置 2 0 B へ発行する (ステップ 8 1 1 0)。副側記憶制御装置 2 0 B では、ライトデータ格納用にキャッシュを確保する (ステップ 8 1 2 0)。キャッシュが確保できたことを正側記憶制御装置 2 0 A へ報告する (ステップ 8 1 3 0)。正側記憶制御装置 2 0 A は報告を受けたらライトデータを転送する (ステップ 8 1 4 0)。正側記憶制御装置 2 0 A は副側記憶制御装置 2 0 B から転送完了報告を受け取る (ステップ 8 1 5 0)。正側記憶制御装置 2 0 A は次のダーティデータがあれば、ステップ 8 1 1 0 ~ ステップ 8 1 5 0 を繰り返す (ステップ 8 1 6 0)。次のダーティデータが無ければ、本処理は終了する。副側記憶制御装置 2 0 B は、非同期にキャッシュからデータを副ボリュームへ格納する (ステップ 8 1 7 0)。

【0 0 3 7】

初期コピー処理は一度全ての差分ビットに対応するデータをコピーする処理である。初期コピーが終了しても、その間ライト要求を受領したら差分ビットマップに '1' が立つ。これらの残コピー処理は、差分ビットマップを先頭から順に検索し、ビットに '1' が立っているところを検出したら、初期コピー処理と同様のコピー処理を行う。

【0 0 3 8】

前記初期コピー中に通常のリード/ライト要求が来る可能性もある。前記ライト要求が来た場合の処理について、図 9 のフローチャートに従って説明する。

ホスト 1 0 からのライト要求を記憶制御装置 2 0 A が受け取る (ステップ 9 0 1 0)。記憶制御装置 2 0 A は、ライト対象データに対応する箇所の差分ビットマップ 2 4 0 P 1 (図 2, 図 1 1 参照) の対象箇所のビットを '1' にする (ステップ 9 0 2 0)。記憶制御装置 2 0 A はライトデータ格納用のキャッシュメモリ領域を確保する (ステップ 9 0 3 0)。記憶制御装置 2 0 A は、ホスト 1 0 からのライトデータを受領してキャッシュに格納する (ステップ 9 0 4 0)。記憶制御装置 2 0 A は、ホスト 1 0 へライト完了報告を返す (ステップ 9 0 5 0)。

記憶制御装置 2 0 A は、ホストからのライト要求とは非同期に正ボリュームへ格納する (ステップ 9 0 6 0)。

【0 0 3 9】

副ボリュームへ前記ライトデータの内容を反映する処理は、差分ビットマップを順次見てビット '1' が検出された場合 (ステップ 9 0 7 0) に、記憶制御装置内ペアであれば (ステップ 9 0 8 0)、ステップ 6 0 3 0 ~ 6 0 6 0, 6 0 8 0 の処理 (ステップ 9 0 9 0)、異記憶制御装置間のペアではステップ 8 0 3 0 ~ 8 0 6 0, 8 1 1 0 ~ 8 1 5 0, 8 1 7 0 の処理を行う (ステップ 9 1 0 0)。次のビットを検索する場合は、ステップ 9 0 7 0 へ進み、なければ終了する (ステップ 9 1 1 0)。

【0 0 4 0】

次に Split 状態の場合について説明する。Split 状態になったら、2 4 0 P 1 と P 2 を切替えて使用する。図 1 1 に示すように、差分ビットマップは正ボリューム側の記憶制御装置内のメモリ 2 2 にもち、1 ペアにつき P 1 と P 2 を持つ。P 1 と P 2 はサイ

ズが同じものとする。Split 状態中に使用する差分ビットマップは正ボリューム側の記憶制御装置 20A 内のメモリ 22 にもつので、副ボリュームへの更新があった場合は、毎回記憶制御装置 20A 内の差分ビットマップをアクセスする。

【0041】

正ボリュームの内容に再同期するリシンク処理 (Resync) の場合、その時点の正ボリュームの内容を副ボリュームにコピーする処理であるため、正副 2 つの差分ビットマップをマージして初期コピーと同様の処理を行う。すなわち、正副どちらか一方のビットに '1' が立っている場合は、全てコピー処理を行う。

【0042】

通常 Split は、初期コピーが終わって正/副ボリューム内容の同期がとれた時点で実行する。これに対し、初期コピー中に Split 要求を受け取った場合でも、ホスト 10 へは即時 Split 完了報告をだし、バックグラウンドで残りのコピーを実行する『高速 Split』がある。図 12 は、本方式の場合の Split 処理を示すフローチャートである。図 12 に示すように、ホスト 10 が『高速 Split』を発行し、記憶制御装置 20A が受領したら (ステップ 12010)、ホスト 10 からのライト要求箇所を格納するための差分ビットマップ 240 を P1 から P2 へ記憶制御装置 20A が切り替える (ステップ 12020)。記憶制御装置 20A は、ペア情報テーブル 220 のペア状態 226 (図 4 参照) を 'Split' に変更する (ステップ 12030)。これで、正/副ボリュームはリード/ライト要求を受け付け可能となる。記憶制御装置 20 は、バックグラウンドで副ボリュームへ未反映のデータをコピーする処理を実行する。副ボリュームへ反映する処理は、差分ビットマップを順次見てビット '1' が検出された場合に (ステップ 12040)、同一記憶制御装置 20 内のペアであれば (ステップ 12050)、ステップ 6020 ~ 6080 と同様の処理を行う (ステップ 12060)。異記憶制御装置 20 間のペアの場合、ステップ 8020 ~ 8070 及びステップ 8110 ~ 8160 の処理を実行する (ステップ 12070)。次のビットを検索する場合は、ステップ 12040 へ進み、なければ処理を終了する (ステップ 12080)。

【0043】

前記高速 Split 状態になったときにホスト 10 から正ボリュームがライト要求を受け取ったときの処理を図 13 に示す。

ホスト 10 から記憶制御装置 20A は正ボリュームへのライト要求を受領する (ステップ 13010)。記憶制御装置 20A は、差分ビットマップ 240 P1 (図 11 参照) のライト対象データに対応する箇所のビットを '1' に設定する (ステップ 13020)。そのビットに対応する箇所のデータがキャッシュにあるか否かの判定を行う (ステップ 13030)。ステップ 13030 でキャッシュになれば正ボリュームからキャッシュにリードする (ステップ 13040)。差分ビットマップ P2 のライト対象旧データに対応する箇所のビットが '1' を検出したらステップ 13060 へ進み、検出しなければステップ 13080 へ進む (ステップ 13050)。「1」が立っている場合は、そのビットに対応するデータがまだ副ボリュームへコピーし終わっていないことを示す。キャッシュ内で副ボリュームのデータ用領域 (異記憶制御装置間のペアの場合は、キャッシュ内で仮想化した副ボリューム用領域) に旧データをコピーする (ステップ 13060)。または、キャッシュにデータがなければ、正副用に 2 回リードしてもよい。コピーに伴い、データが正しいか否かを判定するための冗長情報も副ボリューム用に新規に作成しデータへ付随させる。キャッシュに格納したら、差分ビットを '0' にする (ステップ 13070)。以上のステップ 13030 ~ 13070 の処理により、正ボリュームの旧データがキャッシュ内の副ボリューム用データ領域 (異記憶制御装置間のペアの場合は、キャッシュ内の仮想化した副ボリューム用データ領域) に格納されたことになる。ステップ 13070 の後、または、ステップ 13050 の判断で No の場合、対象となったビットへのライトデータをホスト 10 から受け取り、正ボリューム用のキャッシュ領域に格納する (ステップ 13080)。ホスト 10 へライト完了報告を出す (ステップ 13090)。

【0044】

同一記憶制御装置内のペアであれば（ステップ13100），非同期でキャッシュ上のデータを副ボリュームに格納する（ステップ13110）。異記憶制御装置間のペアの場合，実際のデータが格納されている副ボリュームのある記憶制御装置20Bに，前記キャッシュ上の旧データを格納するため，ステップ8110～8160の処理を実行する（ステップ13120）。

【0045】

次に，前記高速 Split 状態になったときにホスト10から副ボリュームがライト要求を受け取ったときの処理を説明する。フローチャートは図13と同じである。

ステップ13010と同様に，ホスト10から記憶制御装置20Aは仮想化副ボリュームへのライト要求を受領する。ステップ13020と同様に，記憶制御装置20Aは，差分ビットマップ240P1（図11参照）のライト対象データに対応する箇所のビットを‘1’に設定する。ステップ13030と同様に，ビットに対応する箇所のデータが正ボリューム用キャッシュ領域にあるか否かの判定を行う。ステップ13040と同様に，ステップ13030でキャッシュになれば正ボリュームからキャッシュにリードする。ステップ13050と同様に，差分ビットマップP2のライト対象旧データに対応する箇所のビットが‘1’を検出したらステップ13060へ進み，検出しなければステップ13080へ進む。ステップ13060と同様に，キャッシュ内で副ボリュームのデータ用領域（異記憶制御装置間のペアの場合は，キャッシュ内で仮想化した副ボリューム用領域）に旧データをコピーする。または，キャッシュにデータがなければ，正副用に2回リードしてもよい。または，キャッシュからリードする場合，直接キャッシュ内で副ボリュームのデータ用領域（異記憶制御装置間のペアの場合は，キャッシュ内で仮想化した副ボリューム用領域）に旧データをリードしてもよい。そうすることで，リード処理が1回ですむ。コピーに伴い，データが正しいか否かを判定するための冗長情報も副ボリューム用に新規に作成しデータへ付随させてキャッシュに格納する。もし，ライトデータがビットに対応する全データと一致するならば，キャッシュに旧データがなくてもよい。領域だけ確保する。キャッシュに格納したら，差分ビットを‘0’にする（ステップ13070と同様）。対象となったビットへのライトデータを受け取り，副ボリューム用のキャッシュ領域に格納する（ステップ13080と同様）。ホスト10へライト完了報告を出す（ステップ13090と同様）。後はステップ13100～13120と同様である。副ボリュームへのライト要求を出すホストは正ボリュームへライト要求を出すホスト10と別のホストでもよい。

【0046】

次に，ボリュームの再同期について図15のフローチャートに従って説明する。Split 状態になった正ボリュームと副ボリュームはそれぞれリード／ライト要求を処理しているため，内容が異なるボリュームとなっている。再同期とは，その時点の正ボリュームの内容に副ボリュームの内容を合わせることを示す。最初にペア情報のペア状態を変更する（ステップ15010）。差分ビットマップ240P1とP2をマージして，差分ビットマップ240P1に格納する（ステップ15020）。差分ビットマップ240P1の‘1’を検出したら（ステップ15030），初期コピー同様の処理を行う。具体的には，同一記憶制御装置内ペアであれば（ステップ15040），ステップ6020～6080（ステップ15050），異記憶制御装置間ペアの場合はステップ8020～8160となる（ステップ15060）。次のビットを検索する場合は，ステップ15030へ進み，なければ本処理を終了する（ステップ15070）。

【0047】

以上説明したように，コピー元とコピー先のクラスタが異なる場合，実際のコピー先ボリュームをコピー元のクラスタ内にある空きボリュームに仮想的に割り付ける仮想化技術により，同一クラスタ内のペアとしてホストに見せることができる。それにより，クラスタを意識することなく自由に記憶領域のコピーを作成するクラスタ構成記憶装置システムを提供できる。

【0048】

上記にレプリケーション作成処理を説明したが、以下、処理速度優先処理について説明する。処理速度優先処理は、コピーボリュームを作成する処理と通常リード／ライト要求（通常 I／O）を、状況に応じた優先順位に従って処理できるようにするものである。

【0 0 4 9】

前記レプリケーション処理により、コピーに伴う処理を副クラスタの状況にかかわらず、正クラスタから副クラスタに出してしまうが、副クラスタは、通常のホスト I／O を受付処理している。また、副クラスタ内でもコピー機能のエンジンがあり、当該クラスタ内のボリュームのコピーをクラスタ内で作成している。このように、クラスタ毎にコピー機能をもち、ジョブへの要求キューを持ち、ジョブ種類の優先順位に従ってジョブを起動し処理している。

【0 0 5 0】

クラスタ間にまたがったコピー処理を行う場合、クラスタ間またがりで通信する形態を通常のリード／ライト要求形式で行うと、要求を受け取った側は、ホストからの通常 I／O と区別できない。しかし、コピー機能系の処理よりホスト I／O を優先させて実行したい。

【0 0 5 1】

すなわち、異記憶制御装置間のペアの場合、副ボリュームのある記憶制御装置 2 0 B は、正側記憶制御装置 2 0 A から受け取ったライト要求とホスト 1 0 から受け取った通常ライト要求の区別がつかない。従って、記憶制御装置 2 0 B は、全てのライト要求は受け取った順番に処理する。これにより、通常ライト要求の応答時間が遅くなり、性能が低下したように見えてしまうことがある。これらを記憶制御装置 2 0 B が区別して実行順序をスケジュールすることにより、高性能化を実現する。

【0 0 5 2】

以下に示す実施例は、処理優先度をコピー元がコピー先に設定する処理速度優先処理に関するものである。

正側記憶制御装置 2 0 A から受け取ったライト要求とホスト 1 0 から受け取った通常ライト要求を区別する単位として、ボリューム単位を考える。記憶制御装置 2 0 B 内にあるボリュームでレプリケーション作成中のボリュームの場合、記憶制御装置 2 0 A の側で、ペアを組んでいるかどうかを認識している。記憶制御装置 2 0 B のレプリケーション作成中のボリュームへの要求は、レプリケーションするためのコピー処理なのか、通常リード／ライトなのか、ペア状態により記憶制御装置 2 0 A 内で判断できる。

【0 0 5 3】

これを図 1 6 のフローチャートに従って説明する。記憶制御装置 2 0 A 内の要求処理優先度設定プログラム 2 0 2（図 2 参照）により、前記 2 0 A 内のボリュームとレプリケーションを組んでいる記憶制御装置 2 0 B 内のボリュームへの要求を通常リード／ライトと同等に処理すべきか否かの処理優先度を判断する（ステップ 1 6 0 1 0）。記憶制御装置 2 0 A 内で判断した処理優先度情報を記憶制御装置 2 0 B へ渡す（ステップ 1 6 0 2 0）。記憶制御装置 2 0 B は、メモリ 2 2 内にある処理優先度情報ビットマップ 2 5 0（図 2，図 1 4 参照）に前記処理優先度情報を対象としているボリューム番号のビット位置に登録する（ステップ 1 6 0 3 0）。処理優先度情報ビットマップ 2 5 0 は、ボリューム番号毎の、通常リード／ライトと同様にスケジュールして処理するか否かの情報を管理する。例えば、急いで処理したいので、通常リード／ライトと区別なくスケジュールする場合はビット‘1’、通常リード／ライトを優先させたい場合は、‘0’とする。また、通常リード／ライトを優先度 5 として、それ以外に優先度 4～1 で管理してもよい。その際は、1 ボリュームにつき 1 ビットではなく、数ビット用意する。

【0 0 5 4】

レプリケーション作成中である場合、ステップ 8 1 1 0 で、記憶制御装置 2 0 A がリード／ライト要求を記憶制御装置 2 0 B に発行すると（ステップ 1 6 0 4 0）、記憶制御装置 2 0 B は前記要求を受け取り、要求の対象となるボリュームの処理優先度情報を参照する（ステップ 1 6 0 5 0）、処理優先度情報が通常リード／ライト要求と同等にスケジュー

ール可能であれば（ステップ16060），通常リード／ライト処理キュー261（図2参照）に入れ（ステップ16070），キューから要求が選択されたら，ステップ8120のキャッシュエリアの確保を行い，ステップ8130～8150を実行する（ステップ16080）。ステップ16060で処理優先度情報が通常リード／ライト要求より低ければ，通常リード／ライト処理キュー261と別の処理優先度低キュー262（図2参照）に入れ（ステップ16090），選択されたら，ステップ16080の処理を行う。ボリュームの処理優先度情報度はSVP（Service Processor）などの端末，例として管理端末80から保守員により入力してもよい。

通常リード／ライト処理キュー261及び処理優先度低キュー262は，配列またはリスト構造により，発行時刻順に管理されるキュー構造の制御情報である。

【0055】

仮想化したボリューム313（図7参照）に対応する実の副ボリューム312は，ホスト10からは見えていないボリュームになるため，ホスト10からボリューム312に対して要求が発行されることはない。ボリューム312への要求は必ず記憶制御装置20Aを介してしかこない。従って，ボリューム312の要求の処理を急ぐべきか否かを記憶制御装置20A側で判断できる。

【0056】

前記処理優先度低キューのスケジュールについて図17のフローチャートに従って説明する。処理優先度低キュー262は通常リード／ライト処理キュー261より，処理頻度を低くする。処理頻度 α はユーザの設定によって変えられる。

【0057】

前記処理頻度を α とする。通常リード／ライト処理キュー261に未処理な要求があるか判定する（ステップ17010）。あれば，カウンタ $c=0$ を設定する（ステップ17020）。 $\alpha > c$ であれば（ステップ17030），通常リード／ライト処理キュー261の処理を実行する（ステップ17040）。カウンタ c をインクリメントする（ステップ17050）。通常リード／ライト処理キュー261に未処理な要求があるか判定する（ステップ17060）。あれば，ステップ17030へ進む。無ければ，ステップ17100へ進む。ステップ17030で $\alpha \leq c$ となったら，処理優先度低キュー262の未処理キュー数 $\geq n$ ，又は，処理優先度低キュー262の時刻の古い未処理キューが m 時間以上過ぎているか判定する（ステップ17070）。 m ， n は予め決めた数である。ステップ17070で条件に当てはまれば，処理優先度低キュー262の処理を実行する（ステップ17080）。カウンタ c を0にセットし，ステップ17060へ進む（ステップ17090）。ステップ17070で条件に当てはまらなければステップ17090へ進む。

ステップ17010で未処理な要求がないと判定された場合，処理優先度低キュー262の処理を実行し，ステップ17010へ進む（ステップ17110）。

【0058】

また，ステップ16020で記憶制御装置20Aから記憶制御装置20Bへリード／ライト要求を発行する場合，記憶制御装置20Aは，記憶制御装置20Bの状態を意識せずに発行する。例えば，記憶制御装置20Bが通常リード／ライト処理や記憶制御装置20B内のレプリケーション処理によって，記憶制御装置20Bのキャッシュのダーティ量が多くなっている場合がある。ダーティ量が一定基準値より多くなるということは，キャッシュの空き容量が少なくなるということで，レプリケーション処理を一時とめたりして，キャッシュからディスク装置へデータを確保する処理を優先処理してキャッシュの空き領域を増やす必要がある。記憶制御装置20Bはステップ16050で要求の対象となるボリュームの処理優先度を参照すると共に，ダーティ量を参照し，ダーティ量がある一定基準値（あらかじめ設定しておく）を超えていれば，ステップ8120～8150のキャッシュを確保してデータを受領する処理を遅らせることにより，記憶制御装置20Aからの要求の受け入れ量を調整する。記憶制御装置20Aは記憶制御装置20Bから完了報告を受取るまでは次の要求は出さない。

【0 0 5 9】

次に、正クラスタが副クラスタ側にコピー先ボリュームへの処理が通常リード／ライト要求（通常 I／O）に影響がないように処理するか、通常リード／ライト要求（通常 I／O）と同等に処理するかの情報を与える処理速度優先処理の変形例を示す。前述の実施例とは異なり、以下の変形例はコピー先が処理を優先すべきか判断する処理速度優先処理の例である。

【0 0 6 0】

正クラスタ側から副クラスタ側に送信する要求がコピー処理要求（レプリケーションを作成するための要求）である場合は、正クラスタ側（コピー元）が要求命令に本要求がコピー処理要求（レプリケーションを作成するための要求）であることを示す要求種別を付加し副クラスタ側に送信する。これを受けた副クラスタ側（コピー先）が、副クラスタ側は要求種別が付加されている要求は優先度が低い要求と認識し、処理優先度低キュー 2 6 2（図 2 参照）に入れる。これにより、正クラスタ側から受信した要求種別に基づいて、副クラスタ側が処理優先度を定めることができる。この例では、正クラスタ側は、レプリケーションを作成するための要求にその要求がレプリケーションを作成するための要求であることを示す情報を付加して、その要求を送信するが、優先度が低いかどうかを判断するのはあくまで副クラスタ側である。したがって、副クラスタ側でレプリケーションを作成するための要求に対して優先度を低く処理するかどうかを決定できる。

【0 0 6 1】

他の方法としては、正クラスタ側（コピー元）が、コピー処理要求であるという種別ではなく、汎用的に、要求の優先／非優先の情報または優先度 n を命令に付加し、これを受けた副クラスタ側（コピー先）で前記付加された優先度情報を認識して、その優先度を考慮した処理順序のスケジュールを行う。これにより、副クラスタ側が、正クラスタ側から受信した要求の優先／非優先の情報または優先度 n に基づいて、処理優先度を定めることができる。この例の場合も、レプリケーションを作成するための要求に対して優先度を低く処理するかどうかを決定するのは副クラスタ側である。しかし、この例の場合は、一般的な要求に含まれる要求の優先／非優先の情報または優先度 n に基づいて副クラスタ側が判断するので、副クラスタ側で通常の要求の優先処理によりレプリケーションを作成するための要求を優先処理するかどうかを決定できる。

【0 0 6 2】

正クラスタが、副クラスタ側にコピー先ボリュームへの処理が通常 I／O の影響が無い様に処理するか、通常 I／O と同等に処理するかの情報を与え、副クラスタ側で前記情報を参照して、前者の場合は、コピーに伴う処理は優先度を落として処理し、後者の場合は発生時刻順に処理するようなスケジュールする。

【0 0 6 3】

以上説明した処理速度優先処理により、通常業務の I／O を妨げないでレプリケーションを行うことができる。

【0 0 6 4】

以上、本発明の実施例のクラスタ構成記憶装置システムについて説明したが、以下、前記クラスタ構成記憶装置システムの仮想化技術により、クラスタ間のペアを同一クラスタ内のペアとして実現する際の、ユーザの使い方を図 1 8 により説明する。

【0 0 6 5】

最初に、ユーザはコピー先ボリュームを選択する。コピー元ボリューム 3 1 1 に対し、コピー先ボリューム候補である空きボリューム群 3 2 1, 3 2 2, 3 2 3, 3 2 4 は副ボリュームプールとしてユーザに提示される。副ボリュームプールはコピー元ボリュームと異なるクラスタのボリュームも含む。

【0 0 6 6】

ユーザがプールより副ボリュームを選択する。例えば 3 2 2 を管理端末 8 0 から入力することにより選択する。図 1 8 の例では、選択された 3 2 2 はコピー元ボリューム 3 1 1 が属するクラスタ（記憶制御装置 2 0 A）とは異なったクラスタ（記憶制御装置 2 0 B）

に属する。

【0067】

コピー先ボリュームが異なるクラスタのボリュームとなるため、仮想化する副ボリュームをコピー元のクラスタ内の空きボリュームから選択する。ここではボリューム312を仮想副ボリュームとして選択する。このボリュームは実体を持たないため、HDDなどの物理リソースは必要がない。従って、仮想化用の専用のデバイス番号を予め用意し、そこから選択するようにしてもよい。

【0068】

また、この例ではコピー元と同じクラスタに副ボリューム候補があるので、まずそのボリュームから選択できるように、ユーザに提示するとき、候補の中からシステム側で選択して提示してもよい。また、ユーザへのボリューム候補提示や副ボリュームプールからの選択または仮想ボリュームの選択またはその両方を、管理端末80から入力して行う方法やホストコマンドによって行う方法や記憶制御装置側で自動的に行う方法がある。

【0069】

レプリケーション作成プログラム201（図2参照）は、前述のように管理端末から入力された情報、ホストからのホストコマンドに基づいてペア情報テーブル220への情報の登録を行う。あるいは、レプリケーション作成プログラム201は、自動的にペア情報テーブル220への情報の登録を行う。

【0070】

以上、本発明者によってなされた発明を、前記実施例に基づき具体的に説明したが、本発明は、前記実施例に限定されるものではなく、その要旨を逸脱しない範囲において種々変更可能であることは勿論である。

【図面の簡単な説明】

【0071】

【図1】本発明の一実施例を示す計算機システムの構成図である。

【図2】メモリの一実施例の構成図である。

【図3】管理端末（例えば、管理サーバ）の一実施例の構成図である。

【図4】ペア情報テーブルの一例を示す図である。

【図5】レプリケーション作成処理の概要を示すフローチャートの例である。

【図6】同一記憶制御装置内でのレプリケーション作成処理方式を示すフローチャートの例である。

【図7】異記憶制御装置内でのレプリケーション作成処理方式の一実施例を示す図である。

【図8】異記憶制御装置内でのレプリケーション作成処理方式を示すフローチャートである。

【図9】初期コピー中にライト要求が発行された場合の処理を示すフローチャートの例である。

【図10】ボリューム情報テーブルの一例を示す図である。

【図11】差分ビットマップの一例を示す図である。

【図12】Split処理を示すフローチャートの例である。

【図13】高速Split状態の正ボリュームへのライト要求受領時の処理及び副ボリュームへのライト要求受領時の処理のフローチャートの例である。

【図14】処理優先度情報ビットマップの一例を示す図である。

【図15】再同期処理を示すフローチャートの例である。

【図16】処理優先度情報登録処理及び処理優先度情報によるリード／ライト処理のフローチャートの例である。

【図17】処理優先度低キューのスケジュール方式のフローチャートの例である。

【図18】クラスタ間のペアを同一クラスタ内のペアとして実現する際の、ユーザの使い方を説明するための図である。

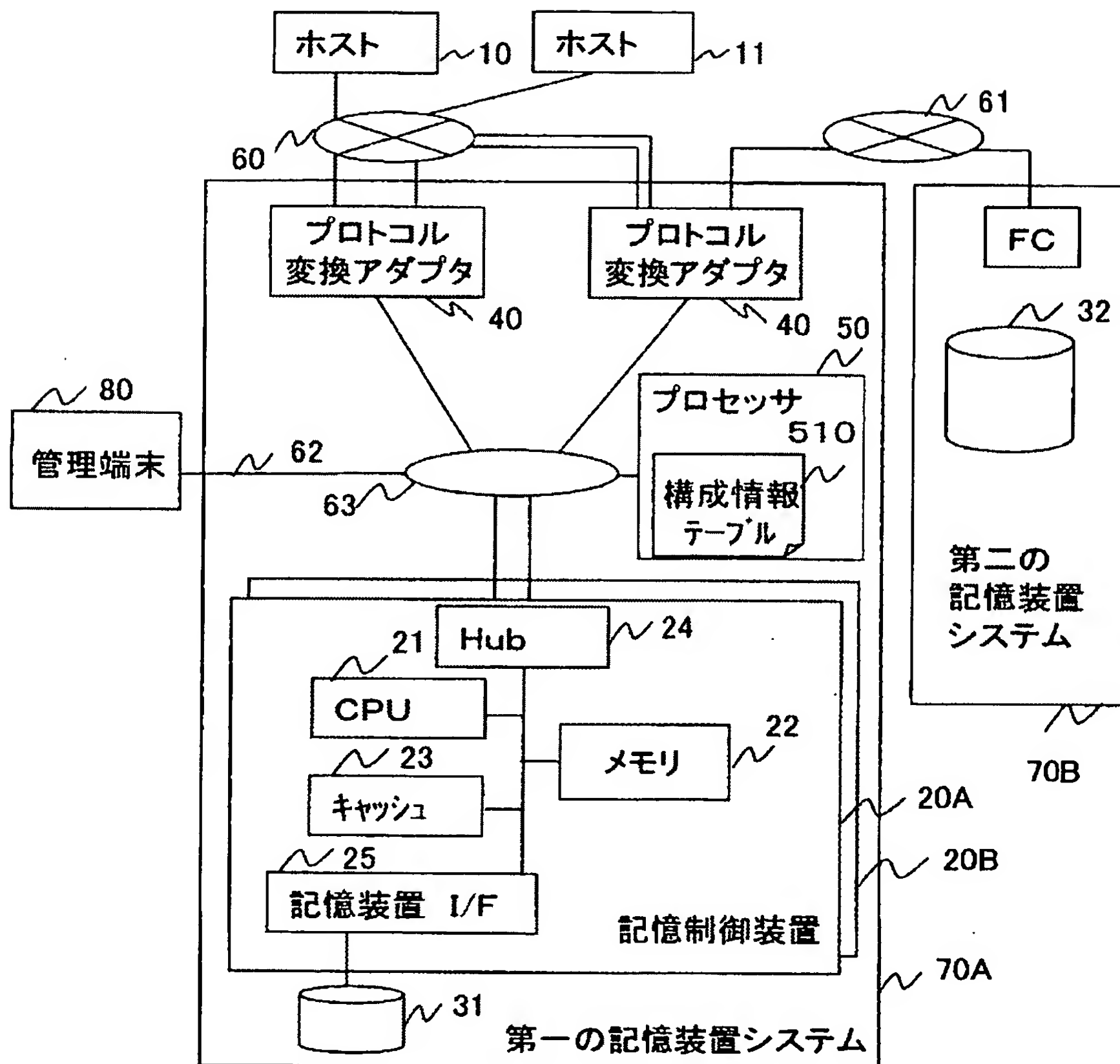
【符号の説明】

【 0 0 7 2 】

1 0 … ホスト, 2 0 … 記憶制御装置, 2 0 2 … 処理優先度設定プログラム, 2 2 0 … ペア
情報管理テーブル, 2 3 0 … ボリューム情報テーブル, 2 4 0 … 差分ビットマップ, 2 5
0 … 処理優先度情報ビットマップ, 3 1 … 記憶装置, 4 0 … プロトコル変換アダプタ, 5
0 … プロセッサ, 6 0, 6 1, 6 2, 6 3 … ネットワーク, 7 0 … 記憶装置システム, 8
0 … 管理端末

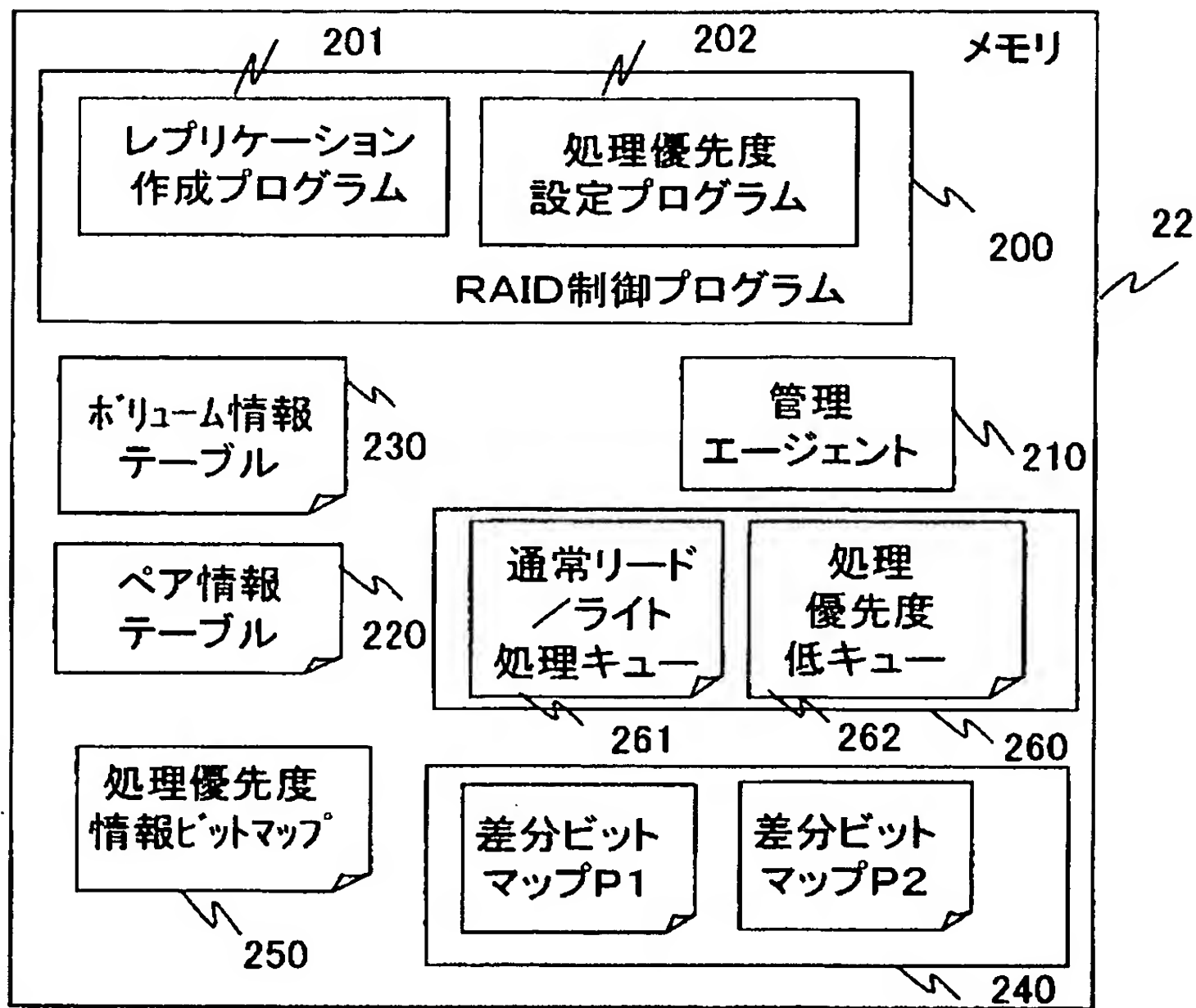
【書類名】 図面
【図 1】

图 1



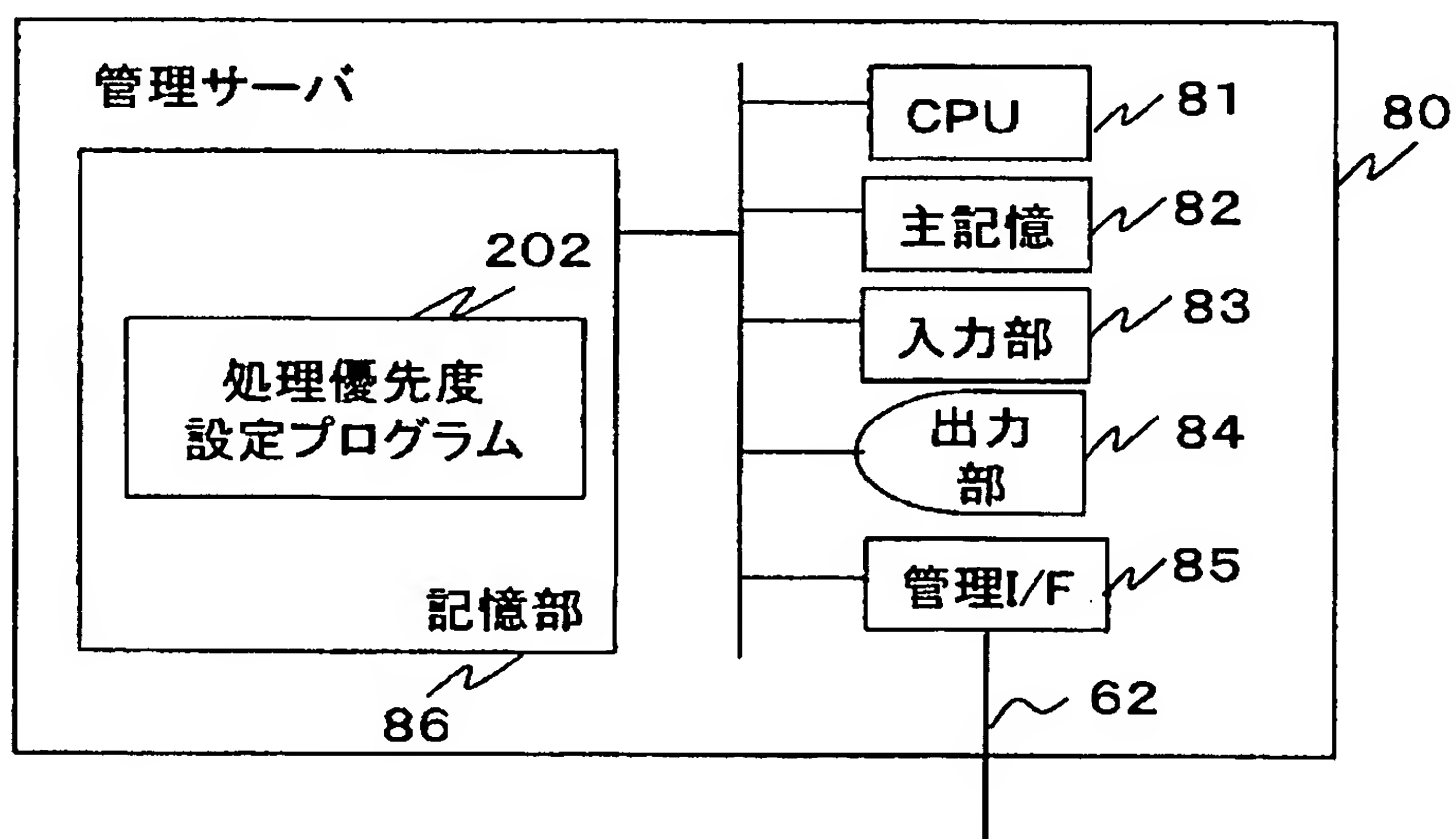
【図 2】

図2



【図 3】

図3



【図 4】

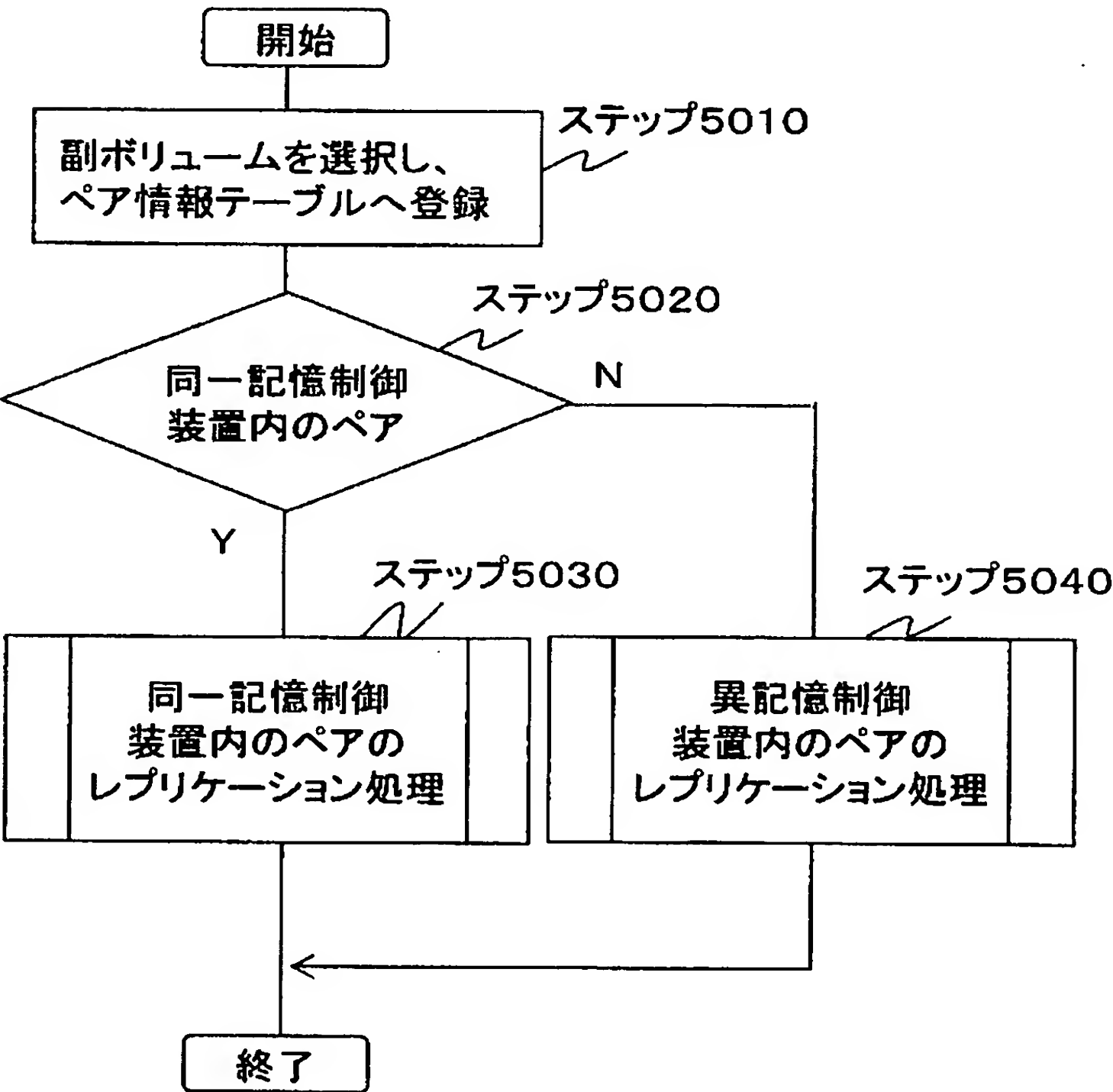
図4

220

ペア 番号	正ボリューム 番号	副ボリューム情報			ペア 状態
		仮想化	実データ格納		
		ボリューム 番号	記憶制御 装置番号	ボリューム 番号	
0	100	200			Pair
1	110	210	2	120	Pair
2	120	220			Split
3	130	230	3	260	Pair
4	140	240			Split
<i>N</i>	<i>N</i>	<i>N</i>	<i>N</i>	<i>N</i>	<i>N</i>
221	222	223	224	225	226

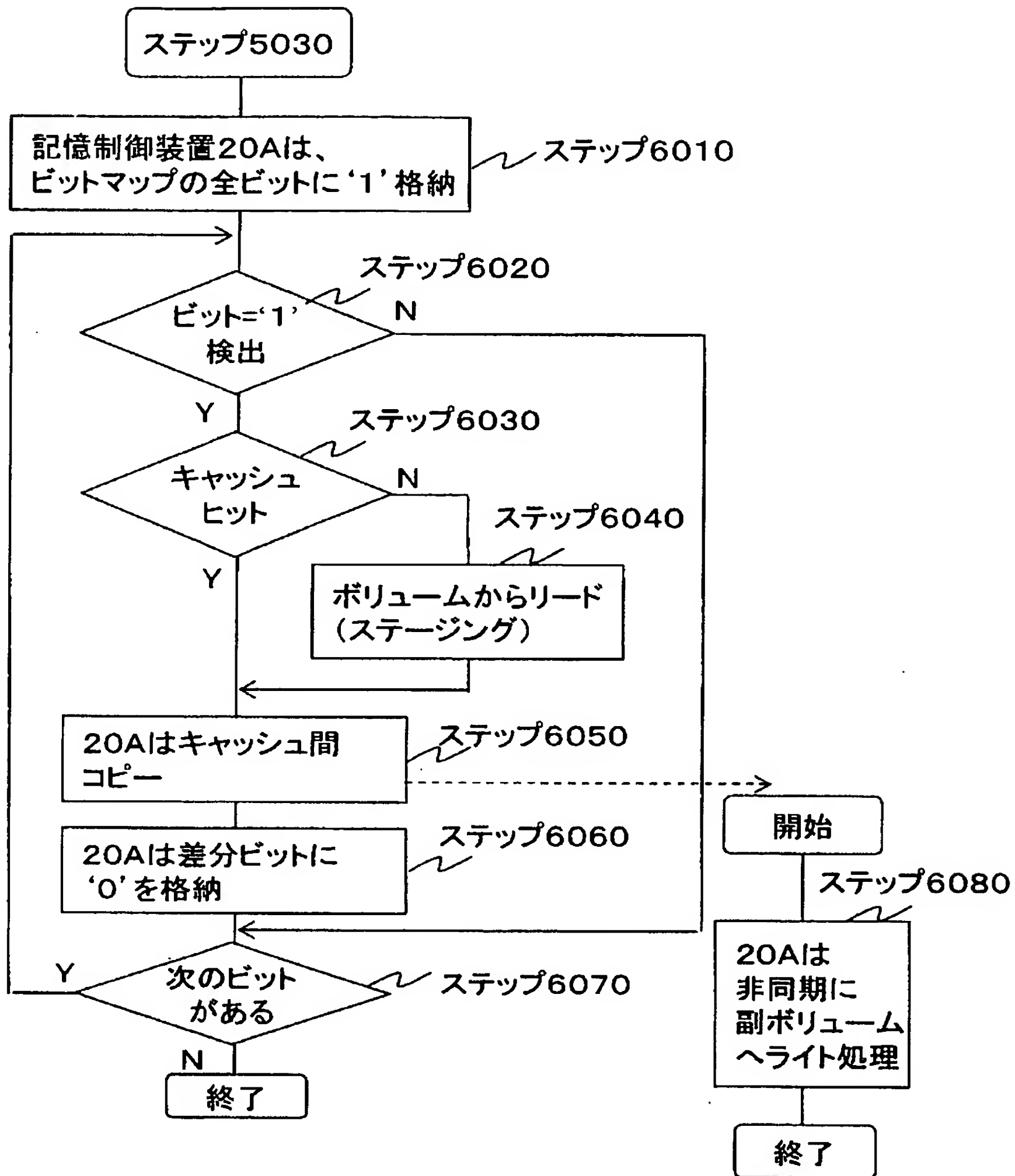
【図 5】

図5



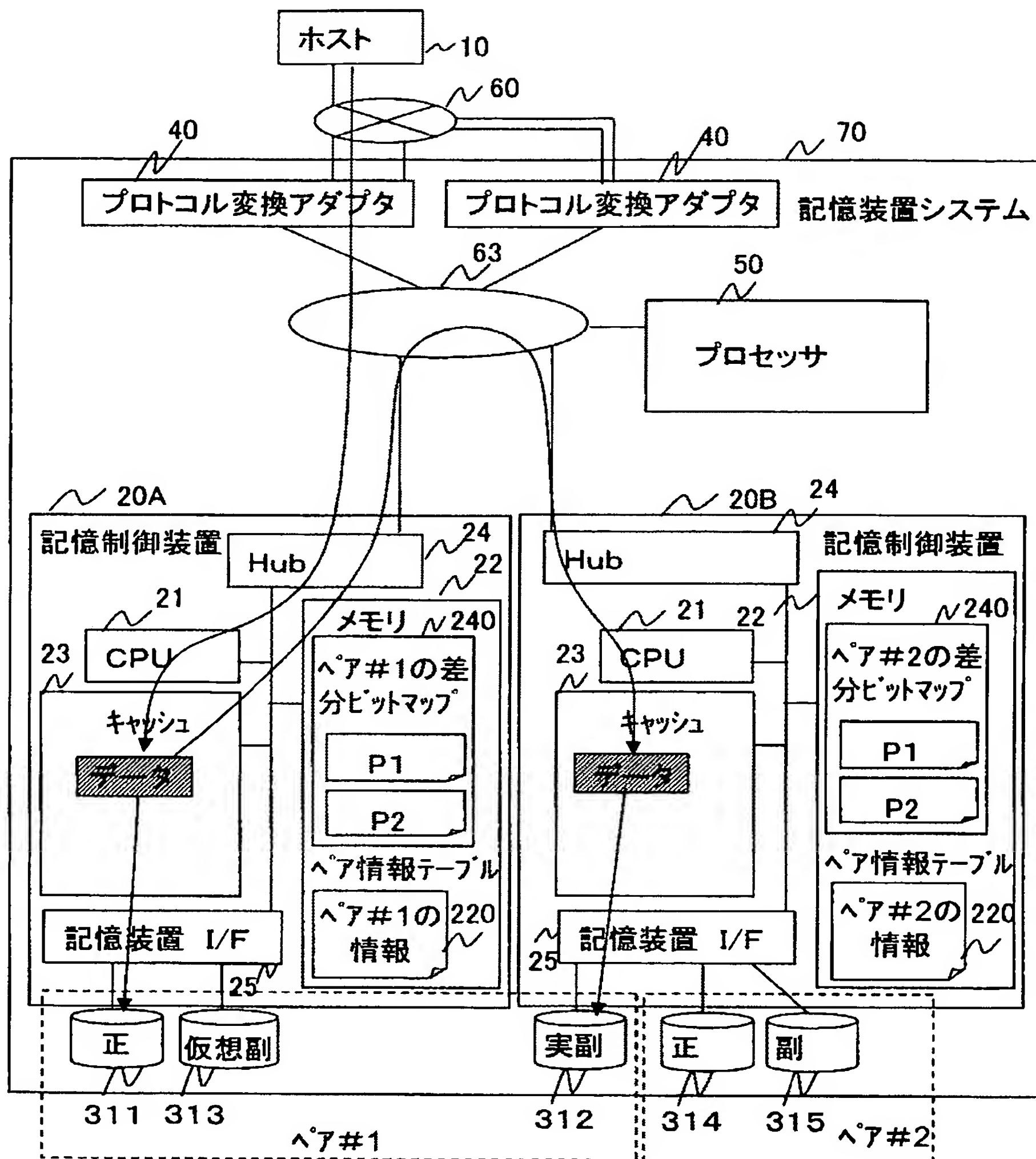
【図 6】

図6



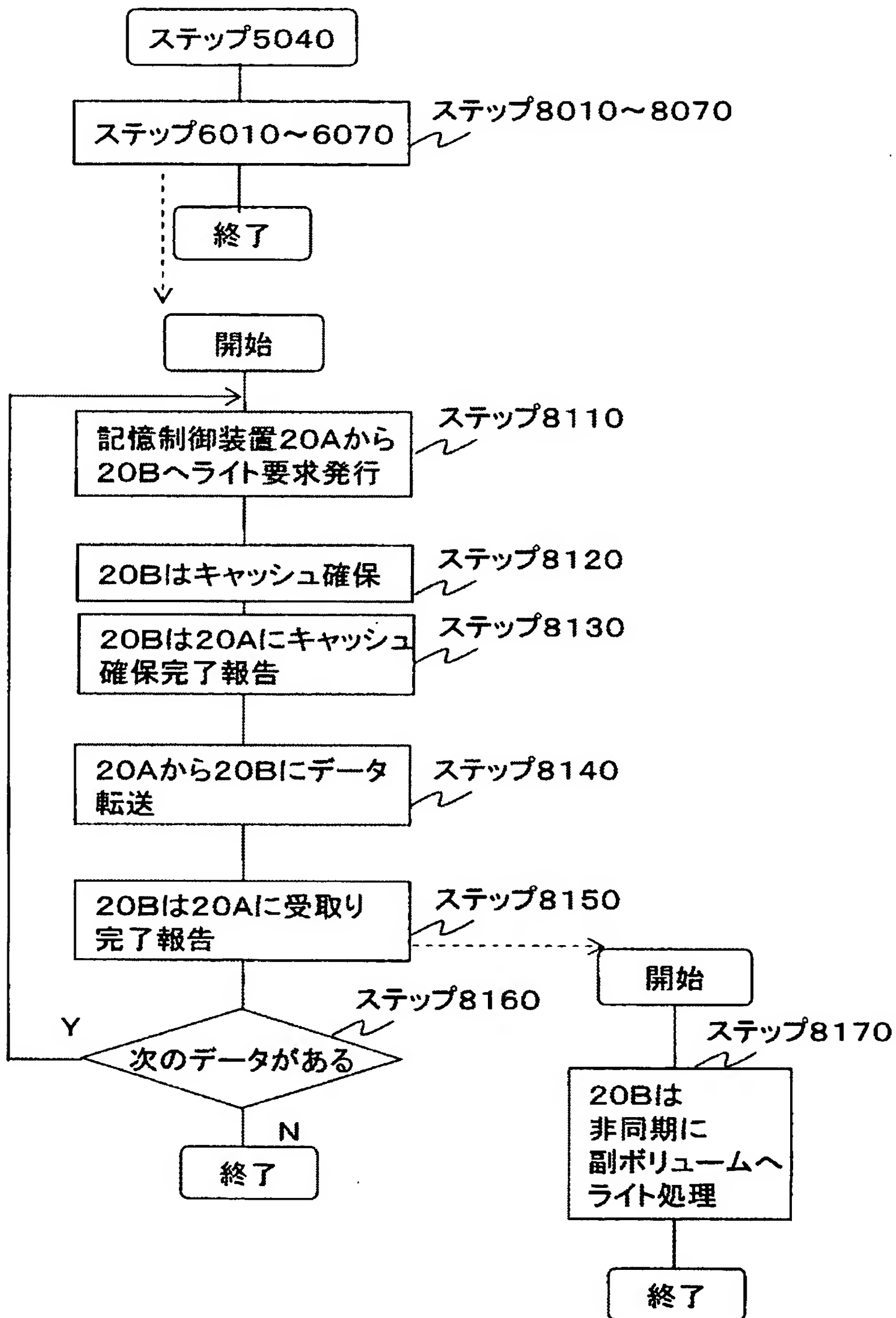
【図 7】

図 7



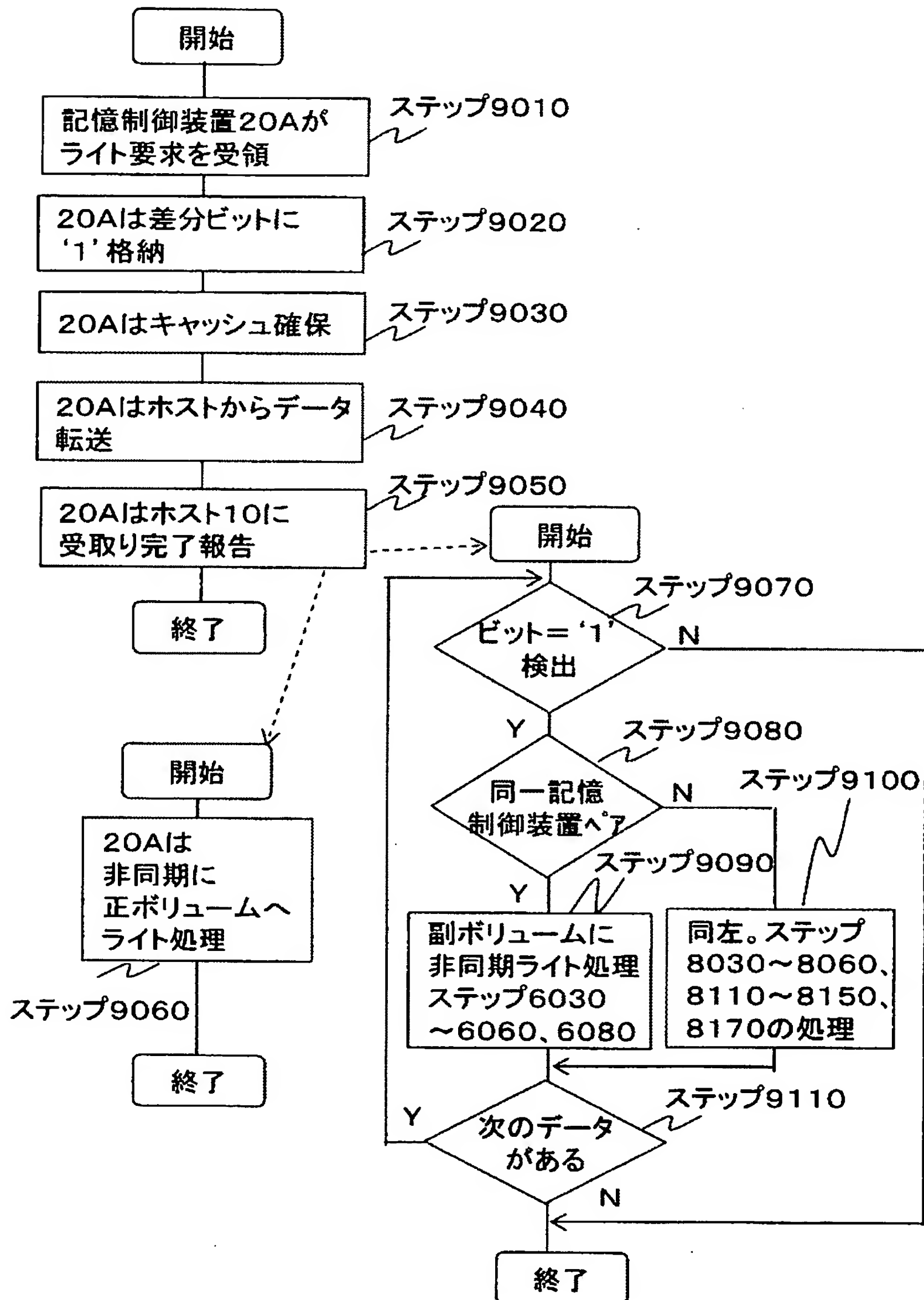
【図 8】

図8



【図 9】

図9



【図 1 0】

図10

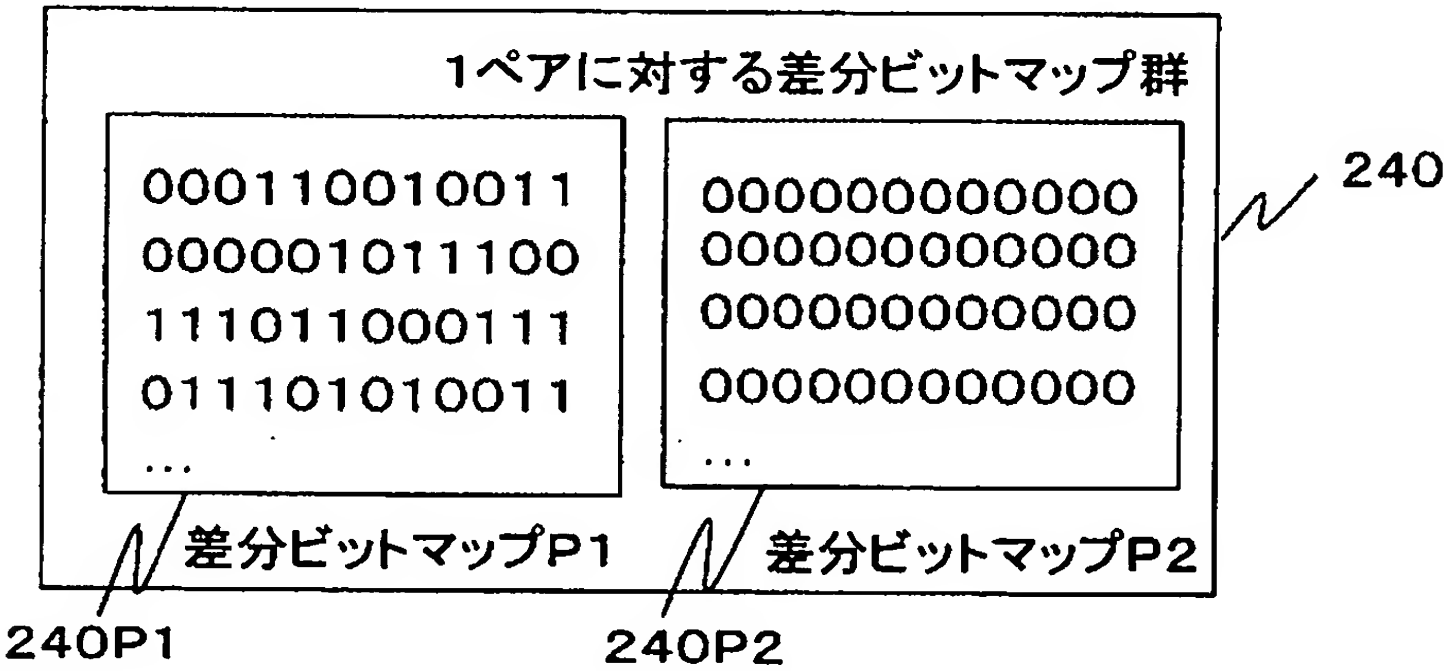
ホリウム 番号	正/副	相手ホリウム情報			ホリウム 使用中
		副ホリウム or仮想化	実ホリウム情報		
			記憶制御 装置番号	ホリウム 番号	
0	正	1024	1	20	使用中
0	正	158	—	—	使用中
0	正	1025	1	426	使用中
1	副		3	3783	使用中

230

231 232 233 234 235 236

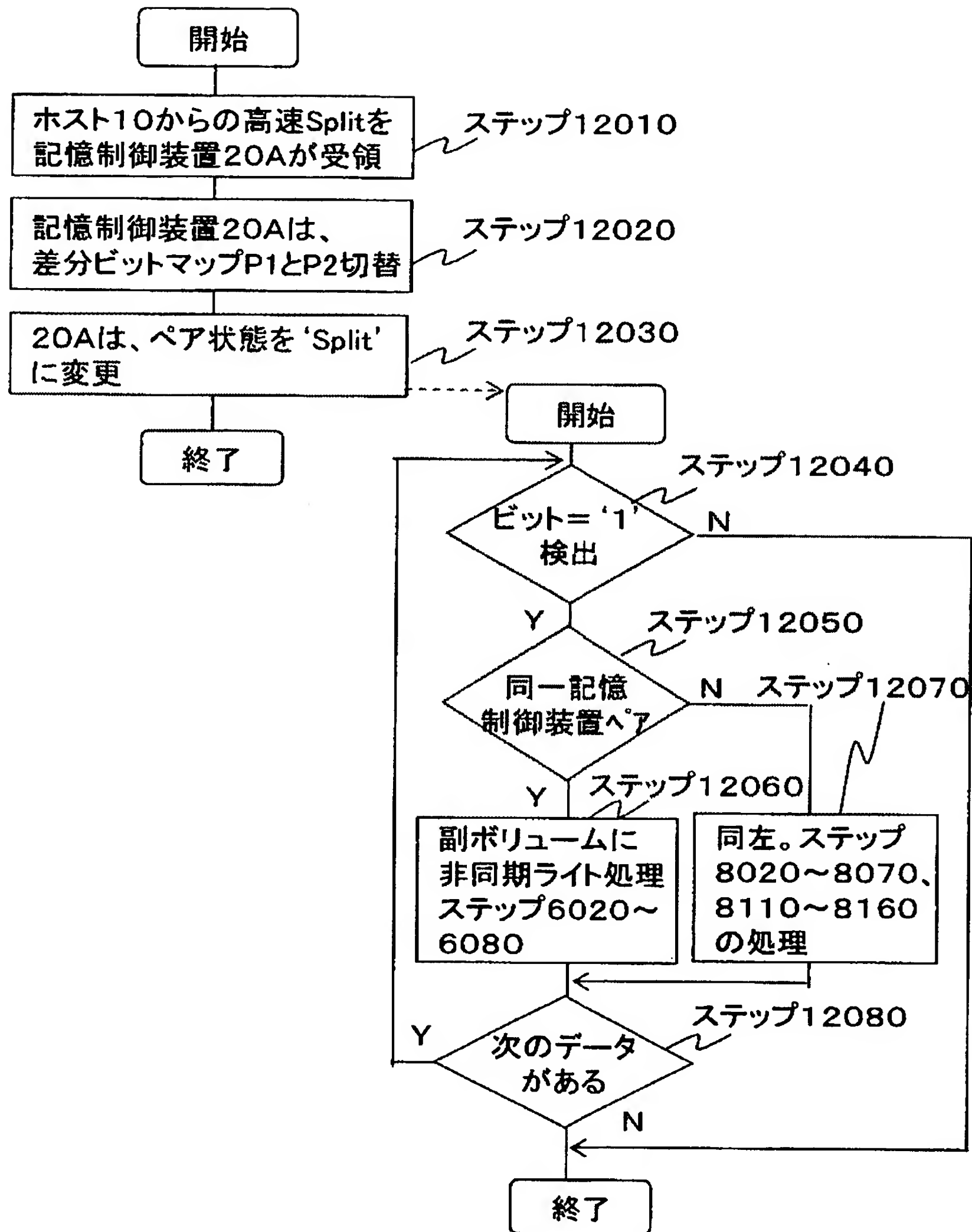
【図 1 1】

図11



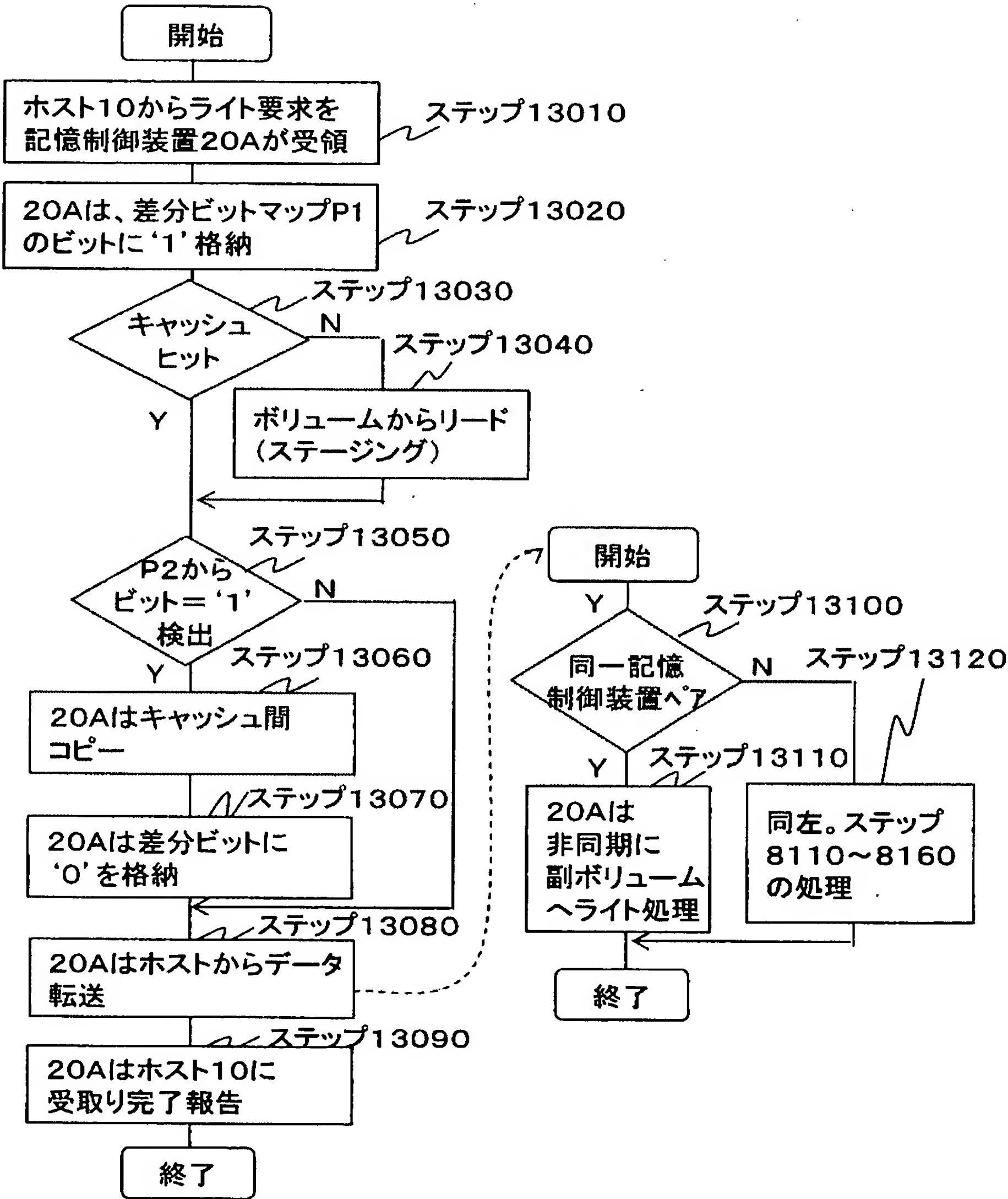
【図 12】

図12



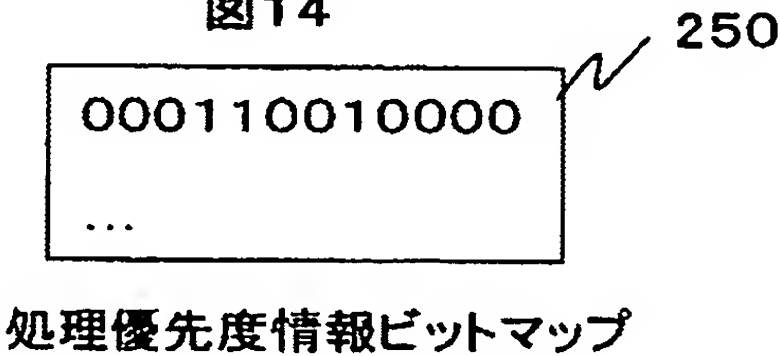
【図13】

図13



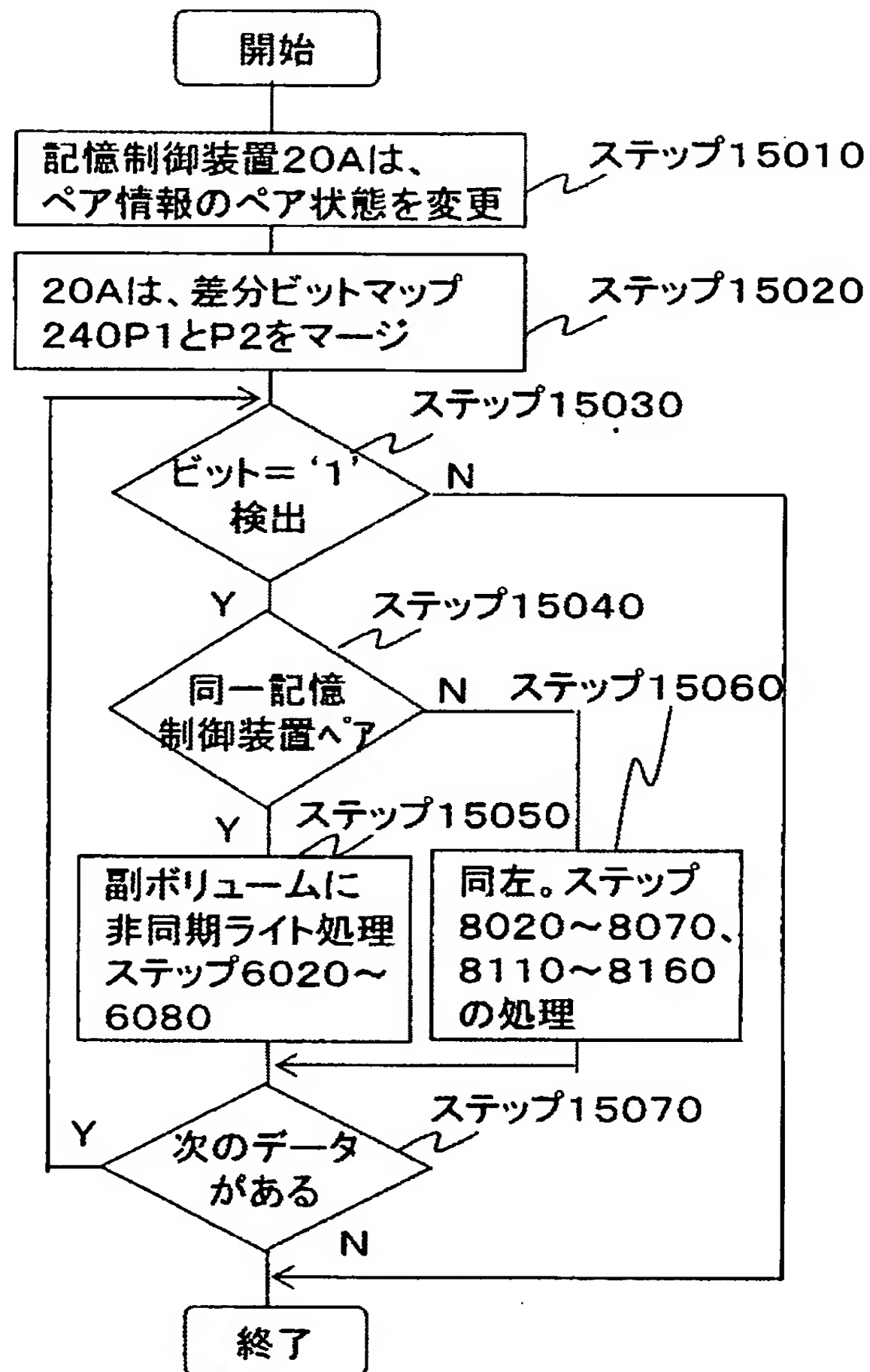
【図14】

図14



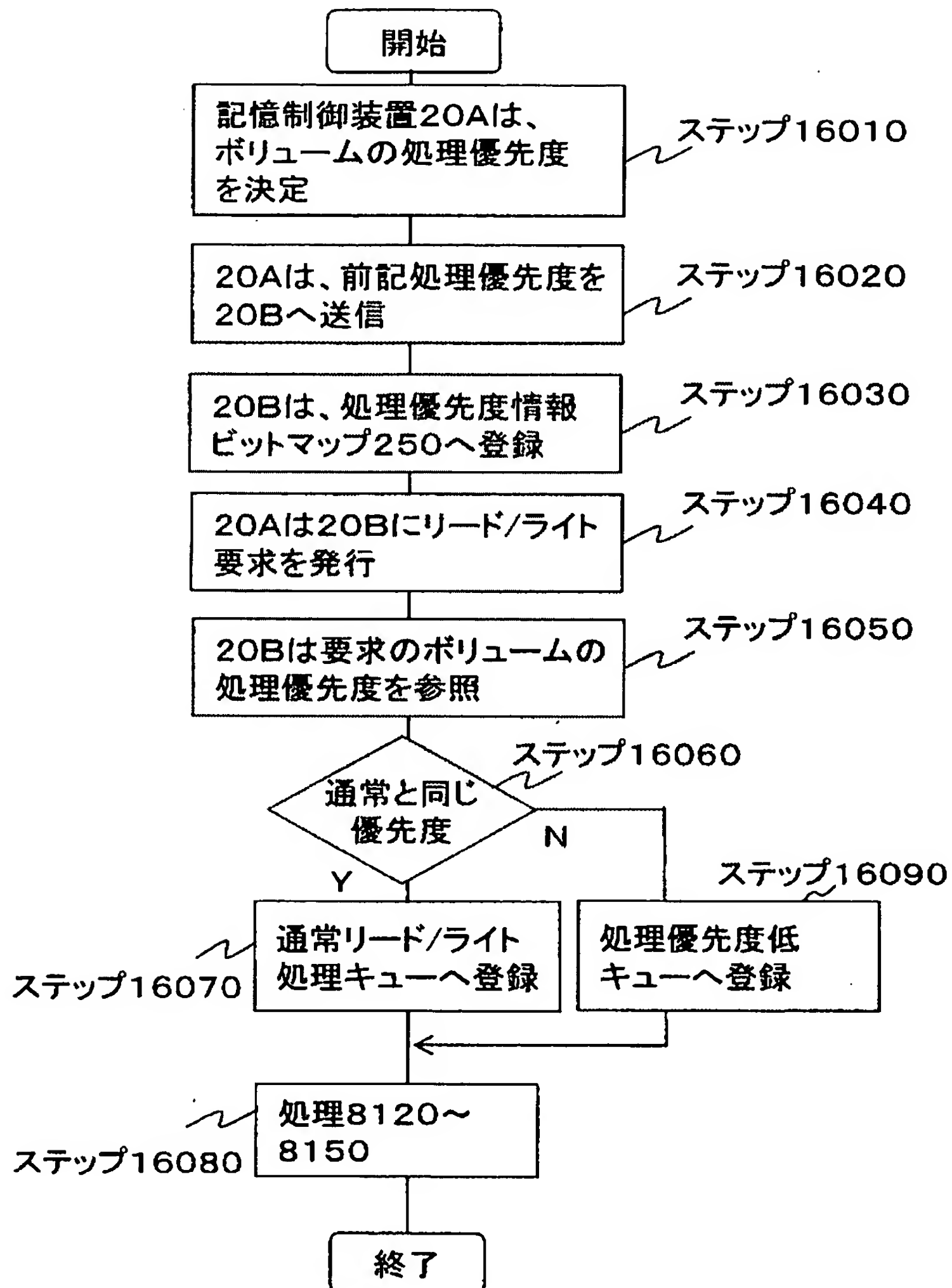
【図 15】

図15



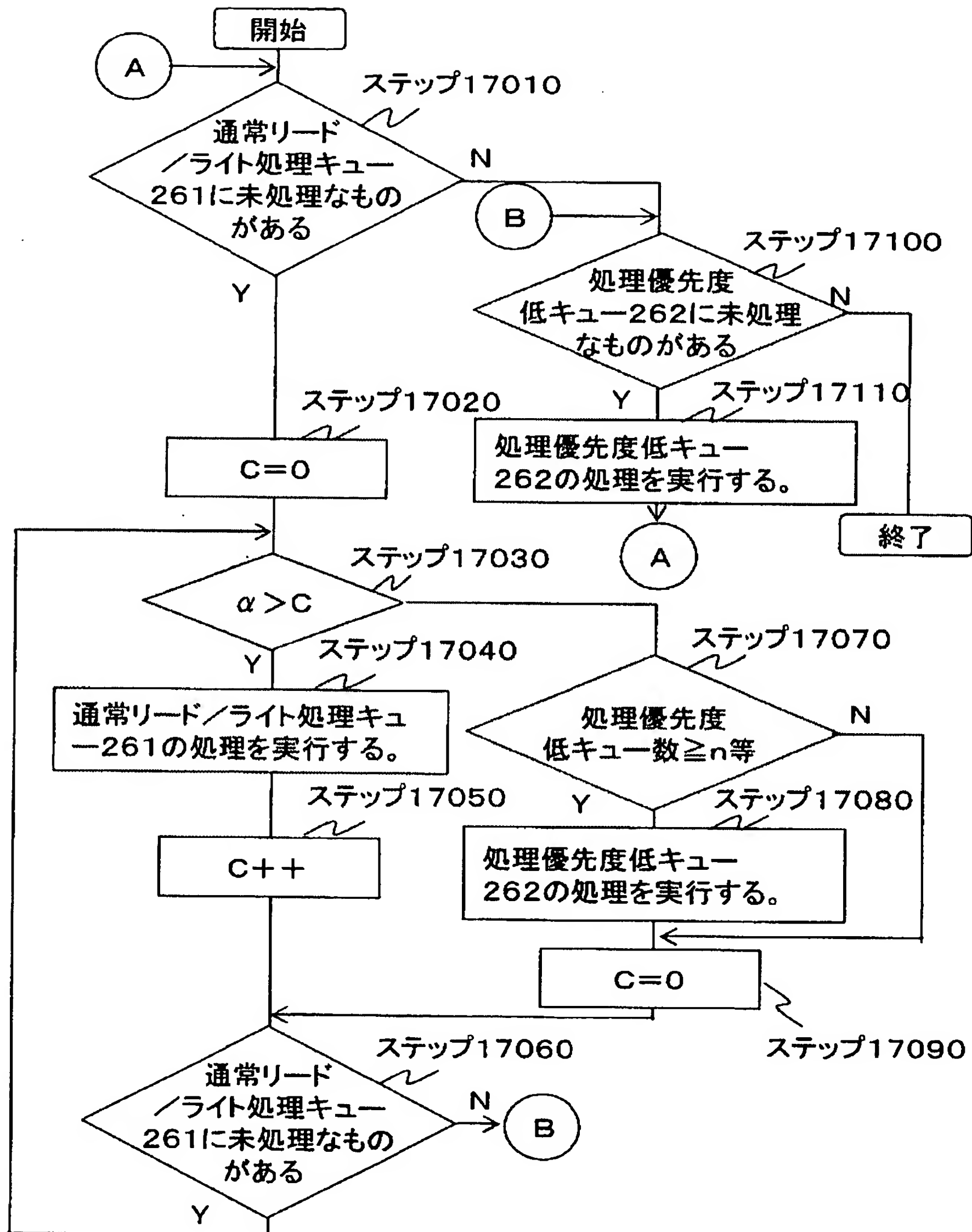
【図 16】

図16



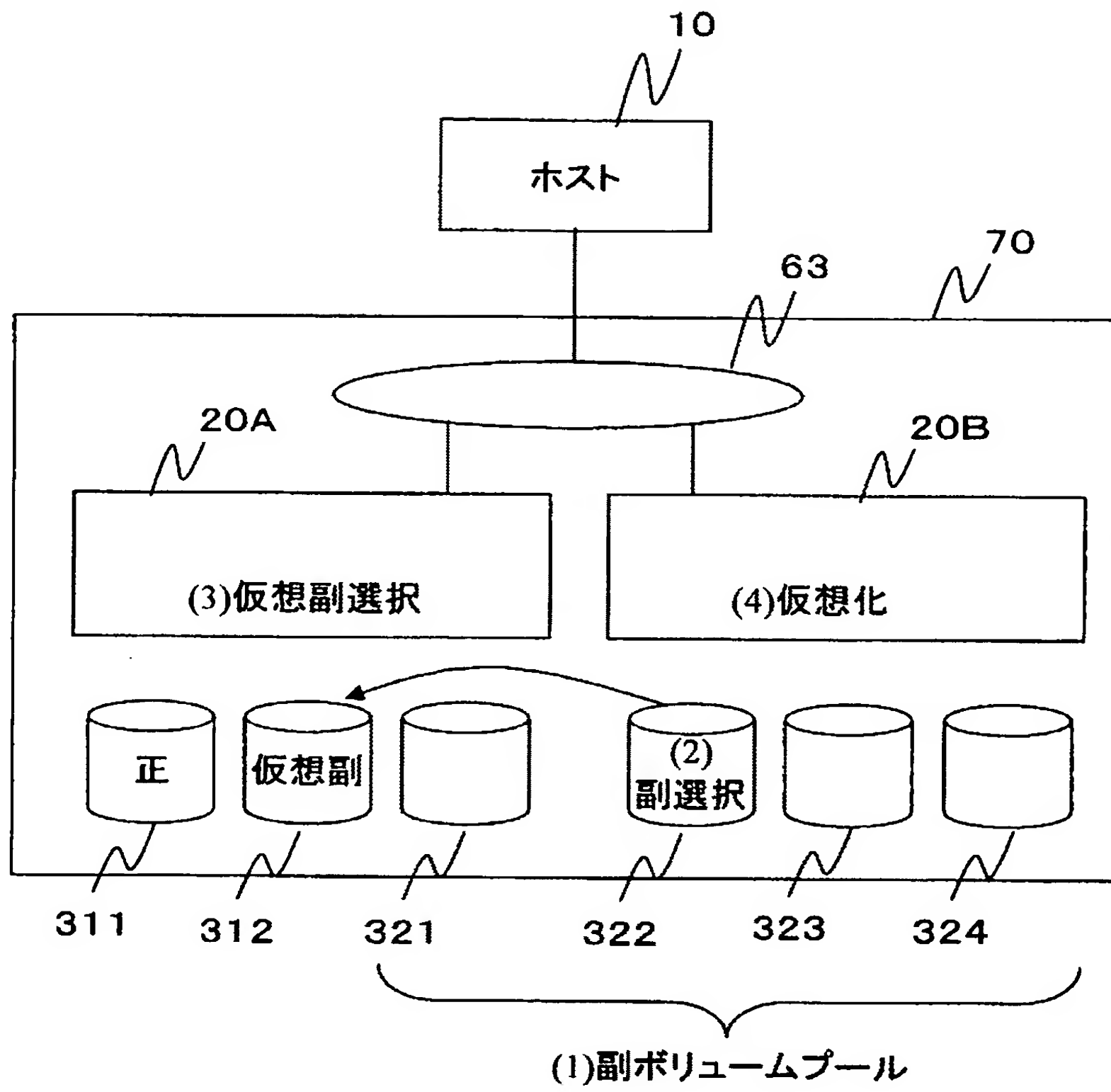
【図 17】

図17



【図 18】

図18



【書類名】 要約書

【要約】

【課題】 複数のディスク装置が接続された制御部を複数個有する記憶装置システムにおいて、異なった制御部に接続されたディスク装置内のボリュームにレプリケーションを作成できる技術を提供する。

【解決手段】 一制御部のレプリケーション作成部は、他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は、ペア情報に、レプリケーション元のボリューム情報と、前記一制御部におけるレプリケーション先のボリューム情報と、前記他の制御部に関する情報と、を登録し、前記ペア情報に基づいて、レプリケーションを作成するための要求を前記他の制御部へ送信する。

【選択図】 図 7

認定・付加情報

特許出願の番号	特願 2 0 0 3 - 4 0 3 8 6 8
受付番号	5 0 3 0 1 9 8 9 4 9 0
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 5 年 1 2 月 4 日

< 認定情報・付加情報 >

【提出日】 平成15年12月 3日

特願 2 0 0 3 - 4 0 3 8 6 8

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所